

# The NCICB User Applications Manual



NATIONAL<sup>®</sup>  
CANCER  
INSTITUTE

Center for  
Bioinformatics



U.S. Department of  
Health and Human Services

## TABLE OF CONTENTS

Introduction.....	1
1.0 Overview of caCORE .....	2
1.1 The NCICB Core Infrastructure.....	2
1.2 Organization of This Manual .....	2
Vocabularies, Ontologies, and Metadata .....	5
2.0 The Enterprise Vocabulary Services.....	6
2.1 The UMLS Metathesaurus .....	6
2.2 The NCI Metathesaurus and the Metaphrase Server.....	7
2.3 Description Logic and the NCI Thesaurus.....	15
2.4 Concept Edit History in the NCI Thesaurus .....	18
2.5 The NCI Ontology Browser .....	20
2.6 The EVS Data Sources.....	27
3.0 The Cancer Data Standards Repository .....	28
3.1 Modeling Metadata: The ISO/IEC 11179 Standard.....	28
3.2 The caDSR Metamodel .....	31
3.3 The CDE Browser .....	37
3.4 The CDE Curation Tool .....	43
3.5 The caDSR Admin Tool .....	52
3.6 Comparison of the caDSR Tools .....	62
Genome Analysis Tools.....	65
4.0 BIOgopher.....	66
4.1 Getting Started .....	66
4.2 Example 1: Formulating Ad Hoc Queries.....	67
4.3 Example 2: Using Local Spreadsheets to Acquire Values.....	72
4.4 Example 3: Merging New Results With A Local Spreadsheet.....	76
4.5 Accessing the EVS Terminologies from BIOgopher.....	78
4.6 The caBIO Data Sources.....	80
5.0 The caARRAY Project .....	84
5.1 The Gene Expression Data Portal .....	85
5.2 caWorkbench.....	104
5.3 The Comparative Genomic Hybridization Viewer: webCGH.....	129
Animal Models and Cancer Images Analysis.....	139
6.0 The Cancer Models Database .....	140
6.1 Searching the caMOD Database .....	140

6.2	Data Submissions.....	146
6.3	The Admin Tool.....	149
7.0	The Cancer Images Database.....	150
7.1	Searching the caIMAGE Database.....	150
7.2	Data Submissions.....	155

## LIST OF FIGURES

Figure 2.2-1 The Metaphrase Welcome Page.....	8
Figure 2.2-2 The Response Page for Basic Search for “cold”.....	9
Figure 2.2-3 The Information Page for a Selected Metathesaurus Concept.....	10
Figure 2.2-4 Related Concepts for the Concept “Glioblastoma”.....	12
Figure 2.2-5 Metaphrase Hyperlinks (in Green) to Entrez Pubmed References. ....	13
Figure 2.2-6 The Advanced Options Menubar. ....	14
Figure 2.2-7 Search Results for “Cold” After De-Selecting “Short”.....	14
Figure 2.3-1 An Earthquake in a Semantic Network of News Stories .....	15
Figure 2.4-1 History Records for the Split Action.....	19
Figure 2.5-1 The NCI Ontology Browser.....	20
Figure 2.5-2 Resetting IE’s Cookiehandling Options.....	21
Figure 2.5-3 Browsing the NCI Thesaurus .....	22
Figure 2.5-4 Selecting a Concept to Examine .....	22
Figure 2.5-5 The Ontology Browser’s Search Panel .....	23
Figure 2.5-6 Advanced Search Options.....	24
Figure 2.5-7 The Display Panel After a Quick Search .....	25
Figure 2.5-8 Concept Details Shown in the Display Panel.....	25
Figure 2.5-9 Customization Options in the Display Panel .....	26
Figure 3.1-1 Representing Data in the ISO/IEC 11179 Model.....	30
Figure 3.1-2 Abstract and Concrete Components of the Data Representation .....	30
Figure 3.1-3 Many-To-One Mappings of Information Elements in the Metadata Model.....	31
Figure 3.2-1 Information Component Infrastructure in the Metamodel .....	32
Figure 3.2-2 Administrative and Organizational Components of the caDSR Metamodel .....	35
Figure 3.2-3 Components in the caDSR Metamodel for Clinical Trials Data.....	36
Figure 3.3-1 The CDE Browser Welcome Page.....	38
Figure 3.3-2 Display of the Currently Selected Context in the Search Pane.....	38
Figure 3.3-3 The Data Element Search pane .....	40
Figure 3.3-4 Constraining Data Elements by Value Domain .....	41
Figure 3.4-1 The CDE Curation Tool’s Navigation Bar.....	43
Figure 3.4-2 The CDE Curation Tool’s Basic Search Window.....	44
Figure 3.4-3 Enabling the Search Results Menubar .....	46
Figure 3.4-4 Pop-up Dialog for Creating and Removing Designations.....	46
Figure 3.4-5 The Details Page for a Data Element .....	47
Figure 3.4-6 The Create Data Element Form.....	50
Figure 3.5-1 The caDSR Admin Tool Welcome Screen .....	52
Figure 3.5-2 Full-Text Search Screen.....	56
Figure 3.5-3 The 11179 Attribute Search Screen .....	57
Figure 3.5-4 A Part of the Results Table for a Data Element Search .....	58
Figure 3.5-5 Details Page for the AGE Data Element.....	59
Figure 3.5-6 Maintenance Screen for Modifying a Data Element.....	60
Figure 3.5-7 Creating a New Data Element Component .....	61
Figure 3.5-8 Accessing EVS Terms and Definitions.....	62
Figure 4.1-1 The BIOgopher Welcome Page. ....	67
Figure 4.2-1 Search Criteria Form.....	68
Figure 4.2-2 Specifying Search Criteria: the Gene Object and its Attributes.....	68

Figure 4.2-3 Form for Specifying Attribute Values.....	69
Figure 4.2-4 Updated Query Panel After Specifying Attribute Values .....	69
Figure 4.2-5 Browse caBIO for scientific names of taxa.....	70
Figure 4.2-6 The updated query panel .....	70
Figure 4.2-7 Initializing a report format .....	71
Figure 4.2-8 Defining the output fields.....	72
Figure 4.2-9 The Results Screen.....	72
Figure 4.3-1 The Spreadsheet Upload Page.....	73
Figure 4.3-2 Acquiring Values From a Spreadsheet.....	74
Figure 4.3-3 Selecting Fields for the Report.....	75
Figure 4.3-4 Report Design for Pathway Information.....	75
Figure 4.3-5 Results Downloaded to an Excel Spreadsheet .....	76
Figure 4.4-1 Using the Merge Icon to Combine Results .....	76
Figure 4.4-2 Removing Fields From the Previous Results .....	77
Figure 4.4-3 The Final Report Format and Resulting Output.....	77
Figure 4.5-1 The Organ Object Tree.....	78
Figure 4.5-2 The EVS Search Interface.....	79
Figure 4.5-3 Using the Search Area to Locate Terms .....	79
Figure 4.5-4 Icons for browsing the EVS navigation tree .....	80
Figure 5.1-1 The GEDP Welcome Page.....	85
Figure 5.1-2 GEDP's Basic Search Form .....	86
Figure 5.1-3 The EVS Navigator Window .....	87
Figure 5.1-4 The GEDP Search Results Page.....	88
Figure 5.1-5 The Experiment Detail Page .....	89
Figure 5.1-6 Experiment Submission Process .....	90
Figure 5.1-7 The Download Templates Dialog .....	91
Figure 5.1-8 Sample Information Template (Affymetrix platform, species human).....	91
Figure 5.1-9 General Experiment Information Form.....	92
Figure 5.1-10 The File Upload Applet.....	93
Figure 5.1-11 Final Data Submission Page .....	94
Figure 5.1-12 Experiment Submission is Complete Page .....	94
Figure 5.1-13 Flow chart for generating pathway diagrams.....	95
Figure 5.1-14 Select Arrays for Analysis Page.....	96
Figure 5.1-15 Preprocessing Options for Affymetrix Data .....	96
Figure 5.1-16 Preprocessing Options for GenePix Data.....	97
Figure 5.1-17 Defining How the Arrays Will be Grouped for Comparison.....	97
Figure 5.1-18 Display Page for Pathways Represented in the Array .....	98
Figure 5.1-19 SVG Diagram for the AKT Signaling Pathway.....	99
Figure 5.1-20 Pathway Summary Report for Affymetrix Data .....	99
Figure 5.1-21 Pathway diagram legend .....	100
Figure 5.1-22 Generating Affymetrix data files .....	101
Figure 5.1-23 GEDP Analysis Tools .....	102
Figure 5.1-24 Experimental Data Available for Affy Cel File Analysis .....	102
Figure 5.1-25 The Affy Cel File Search Page.....	103
Figure 5.1-26 Partial Listing of an Affy Cel Search Results Page .....	103
Figure 5.2-1 Layout of the caWorkbench Graphical Interface .....	106

Figure 5.2-2 The Open, Save, and New Operations .....	107
Figure 5.2-3 The Local and Remote Open File Dialog Boxes .....	108
Figure 5.2-4 Opening a Remote File.....	109
Figure 5.2-5 The Remove and Rename Operations.....	109
Figure 5.2-6 An Example Project Tree .....	110
Figure 5.2-7 Creating a New Marker Panel Set.....	111
Figure 5.2-8 Adding a Probe to a Marker Panel .....	112
Figure 5.2-9 caWorkbench Display Immediately After Loading a Data File.....	113
Figure 5.2-10 Pop-up Menus in the Marker Panel Windows .....	114
Figure 5.2-11 Pop-up Menus in the Phenotype Panel Windows .....	115
Figure 5.2-12 Defining Phenotype Panels .....	115
Figure 5.2-13 The Commands Menu Options .....	116
Figure 5.2-14 Graphical Controls in the Microarray Panel .....	117
Figure 5.2-15 Sample Display in the Microarray Panel .....	118
Figure 5.2-16 An Expression Profile Using Phenotype And Marker Panels.....	119
Figure 5.2-17 Graphical Controls in the Color Mosaic View.....	121
Figure 5.2-18 A Color Mosaic View of a Merged Data Set .....	120
Figure 5.2-19 The Tabular View .....	121
Figure 5.2-20 The Marker Annotations View.....	122
Figure 5.2-21 The caBIO Pathways View .....	123
Figure 5.2-22 Using The Image Viewer .....	124
Figure 5.2-23 Using the Affy Detection Calls as a Filter .....	125
Figure 5.2-24 The Dataset History Window.....	126
Figure 5.2-25 An Example of Hierarchical Clustering.....	127
Figure 5.2-26 The SOM Clusters View window .....	127
Figure 5.3-1 webCGH Overview .....	129
Figure 5.3-2 Selecting Experiments for Analysis.....	130
Figure 5.3-3 Defining Experiment Groups .....	131
Figure 5.3-4 Summary of User-Defined Groups .....	131
Figure 5.3-5 Generating a Line Plot .....	132
Figure 5.3-6 Line Plot Display of Raw Data .....	133
Figure 5.3-7 Suppressing, Displaying, and Highlighting Selected Experiments.....	133
Figure 5.3-8 Drilling Down to a Single Chromosome.....	134
Figure 5.3-9 Zooming in on a Region of Interest .....	135
Figure 5.3-10 The Annotation Plot Tool (top half).....	135
Figure 5.3-11 Genome Annotation Features for Annotation Plots .....	136
Figure 5.3-12 Filters and Annotation Feature Documentation From the UCSC Gene Server ...	137
Figure 5.3-13 Annotation Plot for Chromosome 8 .....	138
Figure 6.1-1 The MMHCC Cancer Models Database .....	141
Figure 6.1-2 An Example using the basic search query form.....	142
Figure 6.1-3 The EVS DataTree .....	142
Figure 6.1-4 Results returned from a simple query .....	143
Figure 6.1-5 The Advanced Search query form.....	144
Figure 6.1-6 General Information Page for a Retrieved Model.....	145
Figure 6.2-1 Welcome Page for New User .....	146
Figure 6.2-2 Welcome Page for a Returning User.....	147

Figure 6.2-3 The Continuing Submission Process Page.....	149
Figure 7.1-1 The caIMAGE home page .....	150
Figure 7.1-2 The Simple Search Query Form.....	151
Figure 7.1-3 The EVS Navigator Window .....	152
Figure 7.1-4 Search Results for Human Colon Cancer Images .....	153
Figure 7.1-5 Exploring an Image with the Viewer .....	154
Figure 7.1-6 The Advanced Search Query Form.....	155
Figure 7.2-1 The Data Submission Welcome Screen for returning users.....	156
Figure 7.2-2 The Submission Form for New Images .....	157

## LIST OF TABLES

Table 2.2-1 Topics Included on a Concept’s Information Page .....	9
Table 2.2-2 A Part Of The Sources Table Summary For The “Common Cold” .....	11
Table 2.4-1 The NCI Thesaurus Concept History Table .....	19
Table 2.6-1 NCI Local Source Vocabularies Included in the Metathesaurus.....	27
Table 3.2-1 Information Components in the caDSR Metamodel .....	33
Table 3.2-2 Attributes of an <i>AdministeredComponent</i> .....	34
Table 3.2-3 Components in the caDSR Metamodel For Clinical Trials Data .....	37
Table 3.4-1 Component-Specific Browsing/Editing Capabilities in the CDE Curation Tool.....	43
Table 3.4-2 Component-Specific Attributes Available for Block Editing .....	51
Table 3.5-1 Additional Search Criteria for Different Components .....	54
Table 3.5-2 Additional Search Criteria Mapped to Specific Components .....	55
Table 3.6-1 Capabilities Served by the caDSR Tools.....	63
Table 4.4-1 Erroneous Results Obtained From Inconsistently Merged Values.....	78
Table 5.1-1 Platform-Specific Files for Experiment Submission .....	90
Table 5.2-1 File Operations in the Main Menubar .....	110
Table 5.2-2 Visualization Tools in the View Window .....	117
Table 5.2-3 The Filtering Panel Toolset .....	125
Table 5.2-4 The Normalization Panel Toolset.....	126



## **Introduction**

## **1.0 OVERVIEW OF caCORE**

### **1.1 The NCICB Core Infrastructure**

The last decade has produced a wealth of genomic information that has just begun to be examined. With this accumulation of bioinformatic data has come a paradigm shift to translational research, and a directive to more quickly advance discoveries in basic research to multifaceted clinical settings and trials. This calls not only for advanced analytic tools and customized data warehouses, but, in addition, for computational environments and software tools that support the development of complex data-mining and information management tasks.

The National Cancer Institute's Center for Bioinformatics (NCICB) has as its mission the goal of bridging these diverse initiatives via a core infrastructure called the caCORE. The collection of NCICB web sites described in this user manual and in the accompanying technical guide provide web-based analysis tools and integrated data repositories, as well as a rich development environment for implementing bioinformatics applications.

As described in this manual, clinical and basic research scientists can find web-based tools for the analysis of genomic and clinical data as well as for the development of clinical trials protocols. For cancer research scientists, these interfaces provide access to:

- the Cancer Bioinformatics Infrastructure Objects (caBIO)
- the Gene Expression Data Portal (GEDP)
- the MAGE database (via the webCGH Tool)
- the Cancers Models Database (caMOD)
- the Cancer Image database (caIMAGE)
- the Cancer Molecular Analysis Project (CMAP)
- the Cancer Genome Anatomy Project (CGAP)

For the clinical researcher, the Cancer Data Standards Repository (caDSR) provides metadata support for developing clinical trials protocols, and the controlled vocabularies available from the Enterprise Vocabulary Services (EVS) provide a semantic integration of the many diverse medical terminologies in use today.

Behind this array of web tools, data repositories, and biomedical informatics services is the "caCORE stack"—a set of core technologies providing the necessary middleware and knowledge infrastructure to serve the cancer research community. In addition to providing software and data repositories, the caCORE serves a critical role in defining standards – in biomedical nomenclature, data modeling, and shared data elements – as well as in the processes whereby these models and elements are developed. A guiding principle throughout all of the NCICB projects is the need to establish and/or adhere to agreed-upon standards of data representation, exchange, and manipulation

This user manual describes the various types of cancer research information and services that NCICB makes available via the web, and includes step-by-step instructions on how to access these resources along with simple examples.

### **1.2 Organization of This Manual**

The applications and interfaces described here fall roughly into three broad functional categories:

- Vocabulary, Ontology, and Metadata management services;
- Genome Analysis tools; and
- Animal Models and Cancer Images resources and analysis tools.

The caCORE Enterprise Vocabulary Services provide a rich set of standardized, controlled vocabularies for the life sciences, along with tools for the development and curation of such vocabularies. The vocabularies and ontologies managed by the EVS span multiple disciplines and domains, including human and mouse pathology, epidemiology, molecular biology, genetics, clinical trials, patient care, and various other biomedical and bioinformatic application areas. Indeed, most of the tools described in this guide leverage the EVS vocabularies in their user interfaces.

The Cancer Data Standards Repository addresses a related but somewhat orthogonal aspect of data representation and exchange; specifically, the need to standardize the terminology, report forms, and protocols implemented in clinical trials. Based on the ISO/IEC 11179 standard for metadata, the caDSR manages the NCI Common Data Elements (CDEs) and provides a registry in an Oracle 8i database for agreed-upon clinical terms and their usage.

In the previous caCORE release (1.0), the EVS and caDSR projects were related but separate efforts. One of the new features of caCORE 2.0 is the interface between these two components: caDSR users can now access the EVS terminologies and definitions and use these as the basis for curating new data elements. This interaction between the two projects is further enhanced by the new EVS feature, “Suggest New Term,” which allows curators to request new terms as needed. The EVS staff reviews such requests and, working with the caDSR curators, creates new terms to enrich the NCI vocabularies as well.

The first section of this manual focuses on the vocabulary, ontology, and metadata services provided by the EVS and caDSR projects. Two web interfaces to the EVS are described in Chapter 2, the [NCI Metathesaurus Browser](#) and the [NCI Ontology Browser](#). Three interfaces to the caDSR are discussed in Chapter 3: the CDE Browser, the CDE Curation Tool, and the caDSR Admin Tool. Together, the EVS and the caDSR projects address many of the representational needs and standardization issues involved in large scale research and the sharing and exchange of scientific data.

The second section of this manual focuses on the genome analysis and data mining tools newly released by the caBIO and caArray projects. NCICB has previously provided genome analysis tools via the CGAP and CMAP web sites. With caCORE release 2.0, several new tools have been introduced to further this field of research.

BIOgopher, a web-based data mining tool that provides an interface to the caBIO objects, is one such application. BIOgopher provides access to the data sources hosted by NCICB, such as the CGAP, CMAP, and GAI (Genetic Annotation Initiative) databases; as well as to many external sources, including the National Center for Biotechnology Information’s (NCBI) UniGene, Homologene, and LocusLink databases; the Distributed Annotation Server (DAS) at UCSC; and BioCarta Pathway data.

The caArray project offers a suite of web-based genome analysis and visualization tools for the exploration of NCI-hosted microarray and SAGE data, as well as a stand-alone desktop tool (caWorkbench) for the analysis of microarray data in private laboratories. The caArray project’s Gene Expression Data Portal (GEDP) is a microarray database based on the MAGE object model

that supports the MIAME (Minimum Information for the Annotation of a Microarray Experiment) standard. The GEDP web interface provides tools to submit new data, search archived data, and correlate such data with metabolic pathways information.

A third tool made available by the caArray project is the Comparative Genomic Hybridization viewer (webCGH). webCGH is a web-based application for visualizing and mining microarray-based CGH data. webCGH enables users to search for CGH experiments in a database, to create persistent user-defined groupings of experimental bioassays, and to generate whole genome plots with zoom capabilities for focusing on chromosomal regions of interest. Chapters 4 and 5 present the BIOgopher and caArray applications.

The final section of this manual focuses on animal models of human cancer, and presents the newly released databases and web interfaces at NCICB for the study and exchange of mouse models and cancer images data.

The Cancer Models Database—developed to support the goals of the Mouse Models of Human Cancer Consortium ([MMHCC](#))—is described in Chapter 6. The MMHCC is a collaborative program designed to derive and characterize mouse models and to generate resources, information, and innovative approaches to the application of mouse models in cancer research. The caMOD database provides web-based tools to the entire cancer research community for browsing, downloading, and submitting mouse models, and provides access to both repositories and individual investigators from whom such models can be obtained.

Chapter 7 concludes this guide with a description of the Cancer Images Database. The caIMAGE tools allow users to query the Cancer Image Server for images submitted by fellow researchers, to retrieve images and annotations and, more generally, to manage and administer data that has been previously submitted.

The caMOD and caIMAGE databases and web tools together implement the infrastructure and tools envisioned and required by the MMHCC research community. Together with the emice web site described in this final section of the guide, these projects define a central hub of the emerging research world of animal models of human cancer.

## **Vocabularies, Ontologies, and Metadata**

## 2.0 THE ENTERPRISE VOCABULARY SERVICES

The NCI Enterprise Vocabulary Services were developed in response to the need for consistent shared vocabularies among the various projects and initiatives at the National Cancer Institute. Controlled vocabularies are important to any application involving electronic data sharing; two areas where the need is perhaps most apparent are clinical trials data collection and reporting and, more generally, data annotation of any kind.

The NCI EVS is a set of services and resources that address NCI's needs for controlled vocabulary. The EVS project is a collaborative effort of the Center for Bioinformatics and the NCI [Office of Communications](#). The *NCI Thesaurus*, which is a biomedical thesaurus created specifically to meet the needs of the NCI, is produced by the NCI EVS project. The EVS project also produces the *NCI Metathesaurus*, which is based on the National Library of Medicine's Unified Medical Language System ([UMLS](#)) Metathesaurus, supplemented with additional cancer-centric vocabulary. In addition, the EVS project provides NCI with licenses for [MedDRA](#), [SNOMED](#), [ICD-O-3](#), and other proprietary vocabularies.

The NCI Thesaurus is a biomedical thesaurus created specifically to meet the needs of the NCI. A critical need served by the EVS is the provision of a well-designed ontology covering cancer science. Such an ontology is required for data annotation, inferencing, and other functions. The data to be annotated might be anything from genomic sequences to case report forms to cancer image data. The Thesaurus covers all of these domains, as it includes vocabularies pertinent to disease, biomedical instrumentation, anatomical structure, and gene/protein information—to mention but a few of the included specialties.

The Thesaurus has recently been ranked by the National Center for Vital Health Statistics as one of the two best biomedical terminologies in the country and has been nominated as a standard by the Consolidated Health Informatics initiative, the health-related component of the eGOV initiative (<http://aspe.hhs.gov/sp/nhii/News/hixs.htm>). The Thesaurus is updated monthly, keeping abreast of developments in cancer science.

The NCI Thesaurus is implemented as a Description Logic vocabulary and, as such, is a self-contained and logically consistent terminology. The second vocabulary service provided by the EVS is the NCI Metathesaurus. The purpose of the NCI Metathesaurus is not to provide unequivocal—or even necessarily consistent—definitions. Like the UMLS itself, the NCI Metathesaurus provides mappings of terms *across* vocabularies.

In previous releases of the caCORE, the EVS web interfaces provided public access only to the NCI Metaphrase server, which hosts the Metathesaurus database. In this release, the interfaces have been extended to provide access to both the NCI DTS server, which hosts the NCI Thesaurus and several other vocabularies, as well as to the NCI Metaphrase server. Both the Metaphrase and DTS servers are licensed by NCI from Apelon, Inc.

### 2.1 The UMLS Metathesaurus

As noted above, the NCI Metathesaurus is based on the UMLS Metathesaurus, supplemented with additional cancer-centric vocabulary. Excellent documentation on the UMLS is available at the UMLS Knowledge Sources web site at:

<http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML>

A brief overview of the UMLS Metathesaurus is included here, but it is strongly recommended that users who wish to gain a deeper understanding refer to the above web site. Only those features of the UMLS Metathesaurus that are relevant to accessing the NCI Metathesaurus are described here.

The UMLS Metathesaurus is a unifying database of concepts that brings together terms occurring in more than 100 different controlled vocabularies used in biomedicine. When adding terms to the Metathesaurus, the UMLS philosophy has been to preserve all of the original meanings, attributes, and relationships defined for those terms in the source vocabularies, and to retain explicit source information as well. In addition, the UMLS editors add basic information about each concept and introduce new associations that help to establish synonymy and other relationships among concepts from different sources.

Given the very large number of related vocabularies incorporated in the Metathesaurus, there are instances where the same concept may be known by many different names, as well as instances where the same names are intended to convey different concepts. To avoid ambiguity, the UMLS employs an elaborate indexing system, the central kingpin of which is the *concept unique identifier* (CUI). Similarly, each unique concept name or string in the Metathesaurus has a *string unique identifier* (SUI).

In cases where the same string is associated with multiple concepts, a numerical tag is appended to that string to render it unique as well as to reflect its multiplicity. In addition, the UMLS Metathesaurus editors may create an alternative name for the concept that is more indicative of its intended interpretation. In these cases, all three names for the concept are preserved.

Several types of relationships are defined in the UMLS Metathesaurus, and four of these are captured by the NCI Metaphrase interface:

Broader (RB)	The related concept has a more general meaning.
Narrower (RN)	The related concept has a more specific meaning.
Synonym (SY)	The two concepts are synonymous.
Other related (RO)	The relation is not specified but is something other than synonymous, narrower, or broader.

The UMLS *Semantic Network* is an independent construct whose purpose is to provide consistent categorization for all concepts contained in the UMLS Metathesaurus and to define a useful set of relationships among these concepts. As of the 2003AB release, the Semantic Network defined a set of 135 basic semantic types or categories, which could be assigned to these concepts, and 54 relationships that could hold among these types.

The major groupings of semantic types include organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. Each UMLS Metathesaurus concept is assigned at least one semantic type and, in some cases, several. In all cases, the most specific semantic type available in the network hierarchy is assigned to the concept.

## **2.2 The NCI Metathesaurus and the Metaphrase Server**

The NCI Metathesaurus includes most of the UMLS Metathesaurus, with certain proprietary vocabularies of necessity excluded. In addition, the NCI Metathesaurus includes terminologies developed at NCI along with external vocabularies licensed by NCI. These additional sources and vocabularies are described in [Section 2.6](#). The NCI Metathesaurus is available through the

Metaphrase interface described below as well as through the Java application programming interface described in the caCORE 2.0 Technical Guide.

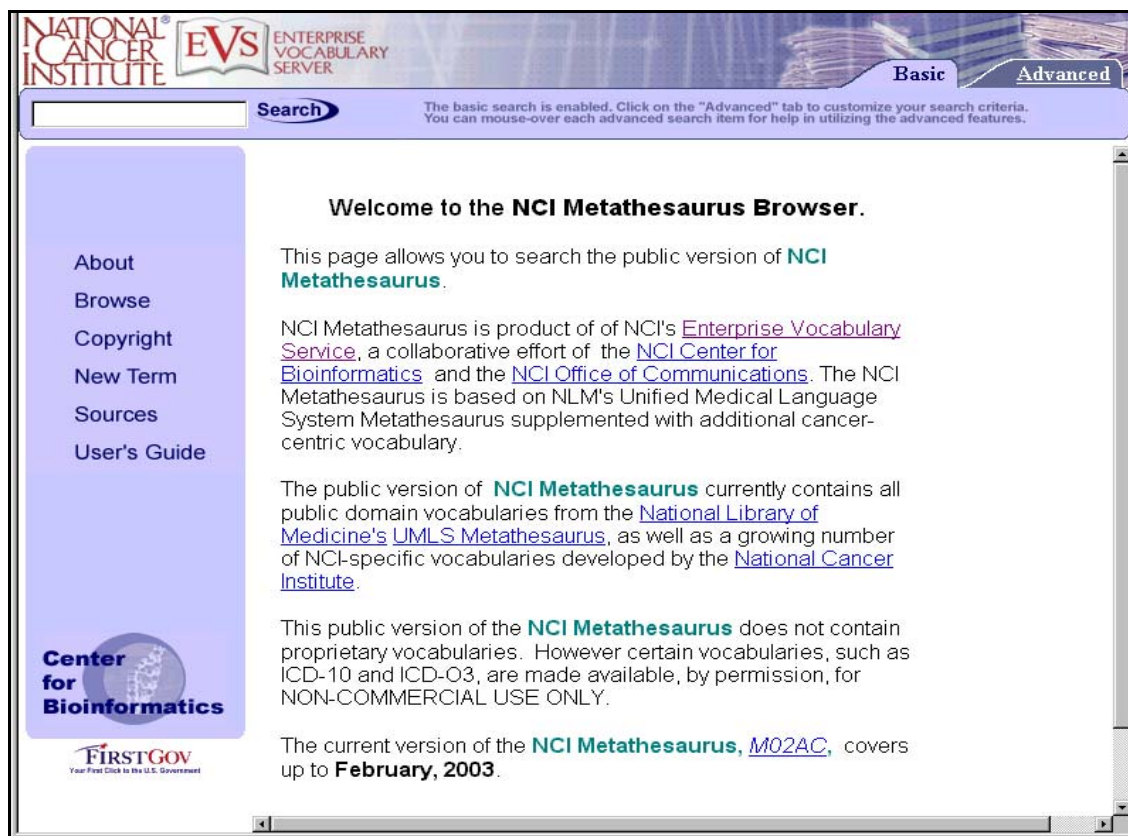


Figure 2.2-1 The Metaphrase Welcome Page

Figure 2.2-1 shows the Welcome Page for the Metaphrase browser. As indicated by the *Basic* and *Advanced* folder tabs at the top of the main display panel, the browser provides two levels of interaction. Immediately below these two tabs is the “search bar.”

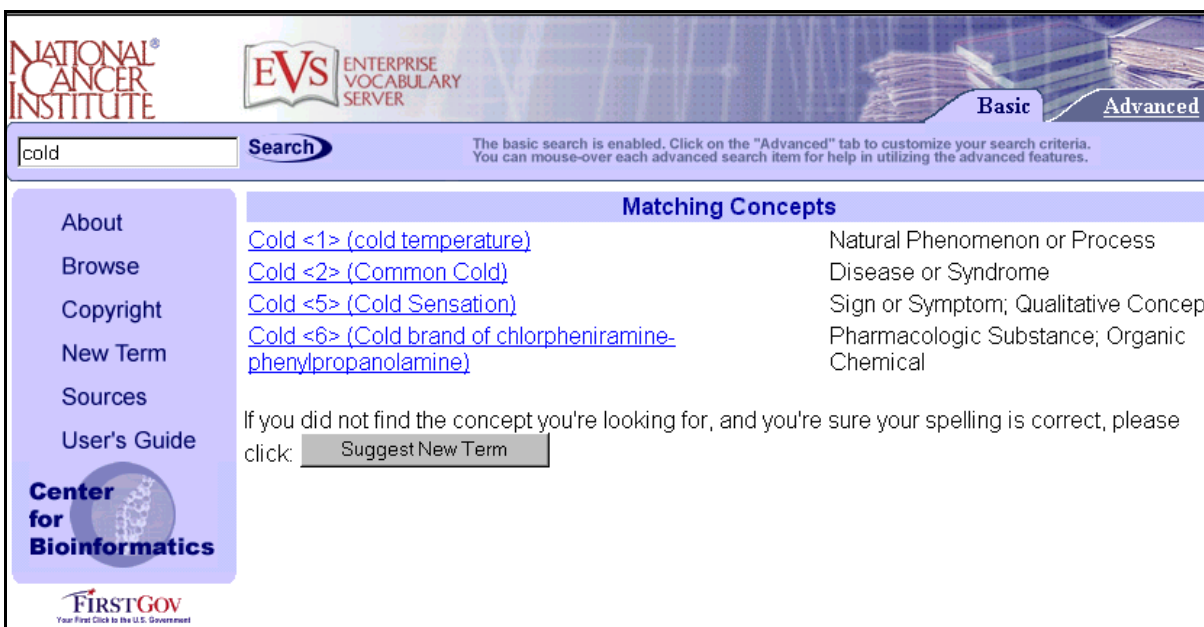
In basic search mode, the user simply enters a keyword or phrase in the search bar’s text box and clicks the **Search** button to submit the query. In response, a page similar to that shown in Figure 2.2-2 appears, showing a list of matching Metathesaurus concepts. In this example, the search term is “cold” and several matches are found.

The first column in this result list shows the concept’s *name(s)* and the second column shows its *semantic type*. There are four concepts whose names had a partial match to the term “cold.” As described in the preceding description of the UMLS, numerical tags have been appended to the concept names to generate unique strings, and the phrases following the tags were added by the UMLS editors to convey the intended meanings. Although each entry has the same concept name “cold,” each one references a different concept. The semantic types listed in the second column also indicate that the four entries correspond to very different concepts.

Clicking on the (selectable) name of a concept in the left column takes the user to the Information Page (Figure 2.2-3) for that concept. This page is a compendium of information culled from all of the sources whose vocabularies include either the concept itself or a known



synonym for the concept. The list of hypertext topics running across the top of the Information Page provides convenient access to the content contained on the page. These topics are summarized in Table 2.2-1.



**Figure 2.2-2 The Response Page for Basic Search for “cold”**

The Information Page shown in Figure 2.2-3 is for the second concept listed in Figure 2.2-2, “Cold <2> (Common Cold).” Immediately below the list of topics is a text bar highlighting the concept’s CUI and its preferred name. “Disease or Syndrome” is the semantic type for the “Common Cold” concept.

**Table 2.2-1 Topics Included on a Concept’s Information Page.**

<u>Topic</u>	<u>Type of Information</u>
Definitions	Indicates the current concept’s meaning
Synonyms	Approximate synonyms
Sources	Sources that contain the concept
Broader Concepts	Parents of the concept
Narrower Concepts	Children of the concept
Related Concepts	Concepts with a semantic relationship to the concept
Medications*	Co-occurring drugs in Medline
Procedures*	Co-occurring clinical procedures in Medline
Laboratory*	Co-occurring lab tests in Medline
Diagnosis*	Co-occurring diagnoses in Medline

\*The last four topics are only provided for concepts that have MeSH (medical subject) headings, and the co-occurrences track only the past three years in Medline.

**NATIONAL CANCER INSTITUTE** **EVS** ENTERPRISE VOCABULARY SERVER

Basic Advanced

Search cold The basic search is enabled. Click on the "Advanced" tab to customize your search criteria. You can mouse-over each advanced search item for help in utilizing the advanced features.

[Concept](#) | [Definitions](#) | [Synonyms](#) | [Sources](#) | [Broader Concepts](#) | [Narrower Concepts](#) | [Related Concepts](#) | [Medications](#) | [Procedures](#) | [Laboratory](#) | [Diagnosis](#) | [Open NCI Hierarchy](#) | [View Hierarchy Location](#)

**C0009443: Common Cold**

Disease or Syndrome

**Common Cold Definitions**

Source	Definition
<a href="#">MSH2002_06_01</a>	A catarrhal disorder of the upper respiratory tract, which may be viral, a mixed infection, or an allergic reaction. It is marked by acute coryza, slight rise in temperature, chilly sensations, and general indisposition. (Dorland, 27th ed)

**Common Cold Synonyms**

- Common Cold
- coryza (acute) <1>
- Cold <2>
- Acute nasopharyngitis [common cold]
- URI (head cold)
- UPPER RESPIRATORY INFECTION VIRAL
- Nasopharyngitis, acute
- Cold (Disease)

**Common Cold Sources**

[AOD2000](#) [CCPSS99](#) [COS95](#) [CSP2002](#) [CST95](#) [ICD10](#) [ICD2002](#) [ICPC93](#) [LCH90](#) [MSH2002\\_06\\_01](#) [MTH](#) [MTHICD9](#)

Center for Bioinformatics

FIRSTGov Your First Click to the U.S. Government

Done Loading Internet

**Figure 2.2-3 The Information Page for a Selected Metathesaurus Concept**

Each definition has a prefix indicating which source provided that definition and the date. In the above example there is only one source definition, from the Medical Subject Headings (MeSH) terminology. Following the list of synonyms is a list of the sources, with each presented as a selectable hyperlink. Clicking on one of these sources produces a page providing further information about the term in that source, including: the term's ID and preferred name [PT] in that vocabulary, synonyms [SY], acronyms [AB], and, if the source is of hierarchical form, the position of that term in the source hierarchy.

The Sources section also provides a link to a tabulated summary of source information, as exemplified by Table 2.2-2. As indicated by this table, it is not uncommon for a concept to have multiple occurrences in a vocabulary due to the multiplicity of names (terms) by which the concept is known. The second column provides the "TTY" code, or term type for each name. The user should refer to the appendices of the [UMLS Knowledge Sources Documentation](#) for the source abbreviations and TTY codes.

Below the list of sources on the Information Page is a pull-down menu box entitled "View Neighborhood." Selecting a source in this box and clicking **OK** produces an expanded list of all *semantically* related ("nearby") concepts in that terminology. Not all vocabularies include semantic relations; these specify additional dependencies beyond the simple inheritances implied

by the concept hierarchies and convey relationships such as “caused by,” “contains,” etc. For vocabularies defining semantic networks of such relations, the concepts included in the “neighborhood” are those removed from the current concept by just one link.

**Table 2.2-2 A Part Of The Sources Table Summary For The “Common Cold”**

Source	Type	Code	Term
AOD2000	DE	0000004799	common cold
CCPSS99	PT	0035226	UPPER RESPIRATORY INFECTION VIRAL
CCPSS99	PT	0059528	COLD
COS95	PT	NOCODE	Cold
CSP2002	PT	3099-9293	common cold
CSP2002	ET	3099-9293	acute coryza
CST95	GT	INFECT	COMMON COLD
ICD10	PT	J00	Acute nasopharyngitis [common cold]
ICD2002	PT	460	Acute nasopharyngitis [common cold]
ICPC93	PT	R74	URI (head cold)
LCH90	PT	U005392	Cold (Disease)
MSH2002_06_01	PM	D003139	Colds, Common
MSH2002_06_01	MH	D003139	Common Cold
MSH2002_06_01	EP	D003139	Cold, Common

## 2.2.1 Navigating Over Related Concepts

Figure 2.2-3 showed only the top section of the information page. Figure 2.2-4 shows the three topics following the Sources information that provide access to additional concepts related to the current selection.

In this second example, the search term is “Glioblastoma.” These related concepts are broken down into Broader Concepts, Narrower Concepts, and Related Concepts.

Not all sources possess hierarchies. Broader and Narrower Concepts are derived from those sources that do contain hierarchy structures. Taken across all sources with hierarchies in which the concept occurs:

- Each concept may have one or more broader concepts whose semantic content is a generalization of the selected concept, and
- A concept may have 0 to many descendants, where each descendant concept is a specialization of the current concept.

Thus, the list of *broader* concepts is the compendium of antecedent concepts from all of the sources that have hierarchies, and the list of *narrower* concepts is the set of all descendant concepts over all such vocabularies.

The list of Related Concepts encompasses a broader and less well-defined set of relations, as it depends on the semantic relations defined in the contributing vocabularies. Some vocabularies, such as the NCI Thesaurus, define very sophisticated and specific relations, such as the fact that a particular bacterium is the etiologic agent of a specific disease. Other sources provide only primitive relations indicating that two concepts depend on one another in unspecified ways.

All of the concepts listed as either Broader Concepts, Narrower Concepts, or Related Concepts are hyperlinked to the corresponding Information Pages for those concepts. Some of these are annotated to indicate the specific relation that is referenced.

For example, the concept “glioblastoma” lists “Common Nervous System Neoplasm” as a broader concept through an *inverse\_isa* relation, meaning that “glioblastoma” *is\_a* (type of) “Common Nervous System Neoplasm.” Similarly, many of the descendant concepts listed as narrower concepts are related to the parent concept through direct *is\_a* relations, e.g., “gliosarcoma” *is\_a* (type of) “glioblastoma.”

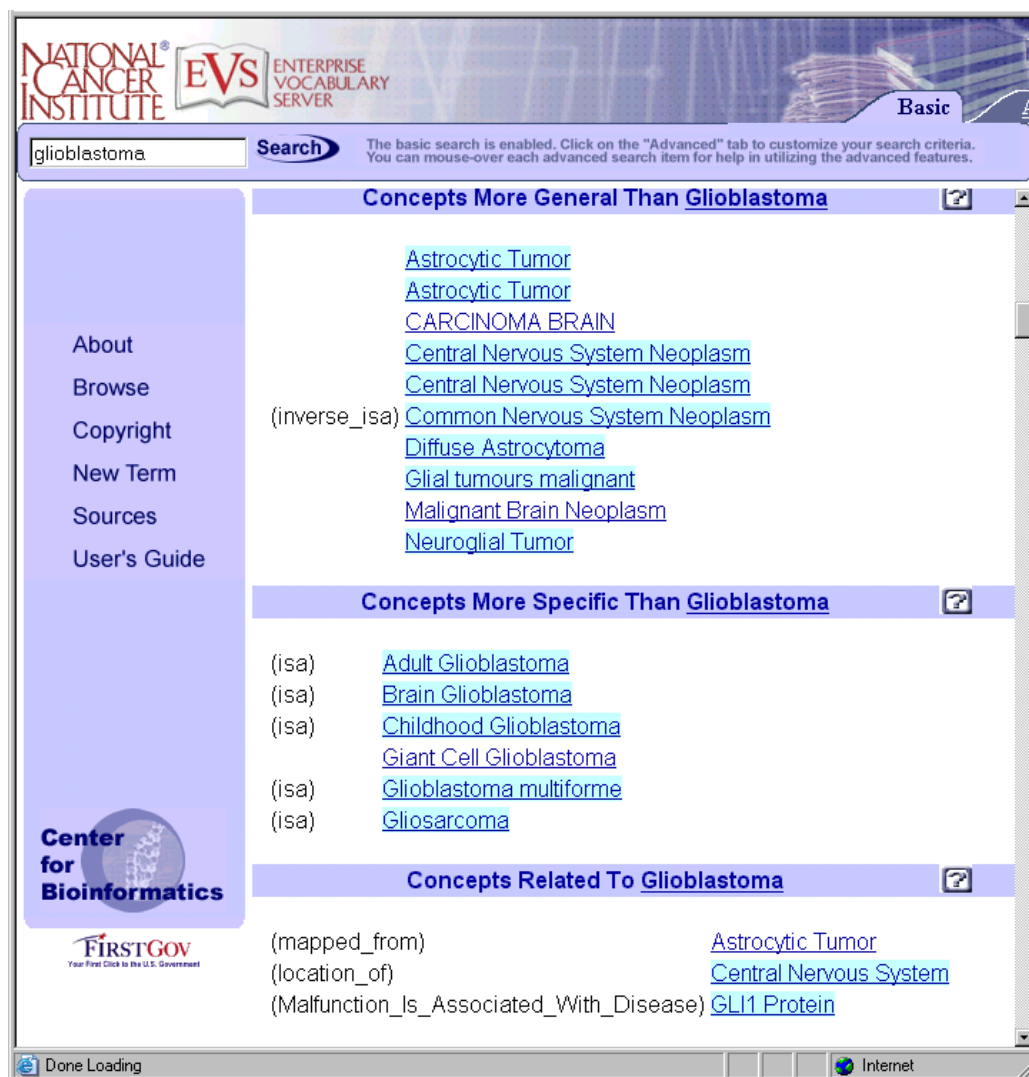


Figure 2.2-4 Related Concepts for the Concept “Glioblastoma”

Some of the related concepts for this example are particularly interesting and include “Astrocytic Tumor,” which is referenced via the *mapped\_from* relation; “Central Nervous System,” which is referenced through the *location\_of* relation; and “GLI1 Protein,” which has the relation *Malfunction\_Is\_Associated\_With\_Disease*. Any related concepts that are contained in one of the NCI local sources are highlighted in blue.

## 2.2.2 MeSH Headings Occurring in the Metathesaurus

In addition to the content described thus far, the Information Pages for concepts that are also MeSH headings provide links to concepts that co-occur with the current concept in Medline. These are grouped into four categories: Medications, Procedures, Laboratory, and Diagnosis. The most commonly co-occurring concepts appear at the top in each category.

Each supplemental concept is preceded by a MedLine hyperlink; clicking on that hyperlink opens a new window connected to the NCBI Entrez browser, which provides a list of clinically relevant articles indexed by Medline. Alternatively, clicking on the concept itself brings up the summary Information Page for that concept provided by Metaphrase.

Figure 2.2-5 for example, shows the procedures associated with the concept “blood cell,” along with the Medline references displayed by the Entrez browser when the link for “stem cell transplantation” is selected.

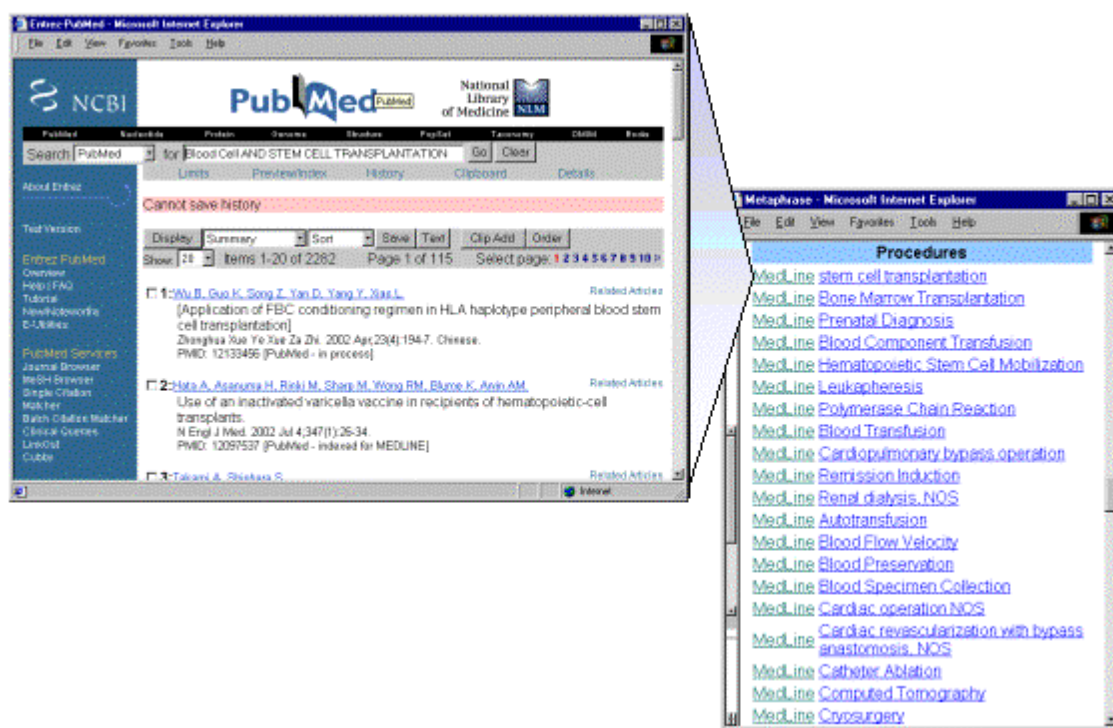


Figure 2.2-5 Metaphrase Hyperlinks (in Green) to Entrez Pubmed References.

## 2.2.3 Advanced Browsing Options

As mentioned at the start of this section, an advanced mode of search is also available. Clicking on the *Advanced* folder tab in Figure 2.2-1 adds the advanced options to the search bar shown in Figure 2.2-6. Starting from the leftmost position, the textbox allows the user to enter search keywords. In the simplest use of this interface, the user merely enters the desired text and presses the **Search** button, as in the Basic interface.

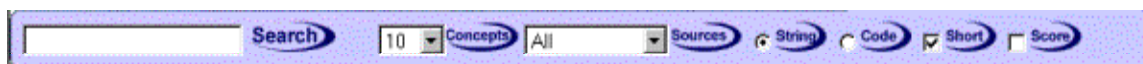


Figure 2.2-6 The Advanced Options Menubar.

The remaining buttons allow the user to:

- Limit the number of matching concepts returned in the results list (**Concepts**);
- Restrict the search to a selected vocabulary (**Sources**);
- Search by string matching or by code (**String, Code**); and
- Limit the number of lexical matches (**Short, Score**).

By default, the maximum number of concepts returned on a single query is 10, and only the highest-quality lexical matches are shown. The drop-down number menu to the right of the textbox allows users to specify that {1, 10, or 100} matches should be returned. The next option allows the user to specify that only selected source vocabularies should be used. Unless explicitly reset, the sources used in concept matching will include all of the vocabularies in the Metathesaurus.

It is also possible to control how the matching will be performed. By default, the user's keywords are matched to concept names. In cases where one knows the concept ID, however, the pair of **string** and **code** radio buttons allow users to override the default.

The last two options control the way in which the *lexical* matching is performed. Lexical matching is based on finding shared “significant” *lexemes* (words or word bases) occurring in both the stored concept names and the words entered in the search box. For example, “degenerative joint disease” is lexically related to “Joints, Knee,” since they share the lexeme “joint.” The NCI Metaphrase server suggests authoritative terms for a given string by calculating a list of the most lexically related terms.



Figure 2.2-7 Search Results for “Cold” After De-Selecting “Short”

The default setting uses a short list of good lexical matches, and can be explicitly selected by checking the box labeled “Short.” The search results for “cold” after de-selecting this box are

shown in Figure 2.2-7. The “Score” option should be selected when complete lexical matching is used, as this option will order the results according to the quality of lexical matches.

## 2.2.4 Additional Options in the Metaphrase Interface

The left-hand sidebar of the Metaphrase pages always lists six additional options:

- *About* provides a quick synopsis of the NCI Metathesaurus and the Metaphrase server.
- *Browse* opens a new window for browsing terms in the NCI Thesaurus vocabulary tree. Initially, the tree contains only a single folder icon labeled “NCI Thesaurus.” Double-clicking on the folder icons in this window will expand that branch of the tree.
- *Copyright* states the copyright and license restrictions for using Metaphrase
- *New Term* provides a form for requesting that the EVS developer team add a new concept or term to one of the Metathesaurus vocabularies. This option is also available at the bottom of the initial search results page.
- *Sources* links to a summary table of 58 sources included in the NCI Metathesaurus.
- *User’s Guide* is a detailed step-by-step how-to guide for using Metaphrase.

## 2.3 Description Logic and the NCI Thesaurus

The encoding of concept-based vocabularies, particularly those that capture a rich body of relationships and features, is an important branch of knowledge representation. One of the most common approaches to knowledge representation in the 1970s was *frame-based* representations.

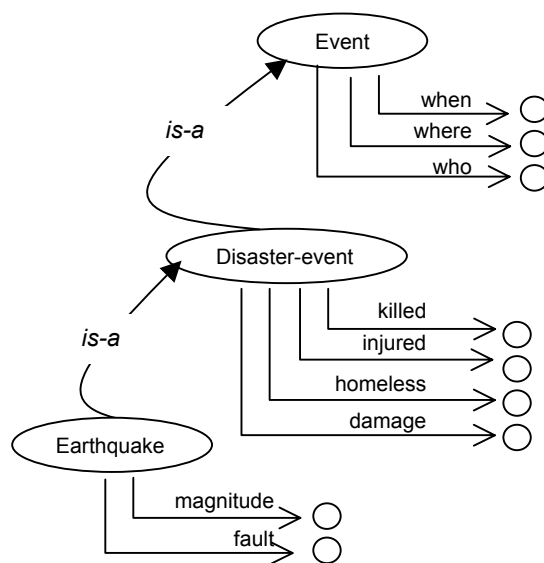


Figure 2.3-1 An Earthquake in a Semantic Network of News Stories

The basic idea of a frame is that the objects in our world fall into natural classes, and that all members of these classes share certain properties or attributes, called *slots*. For example, all dogs have four legs, a tail (or vestige of one), whiskers, etc. Thus, when we encounter a new dog, we already have a “frame of reference” and some expectations about the properties and behaviors of these entities.



In a seminal paper by Marvin Minsky published in 1975, he placed the frame representation paradigm in the context of a *semantic network* of nodes, attributes, and relations. Figure 2.3-1 shows a simple frame-based representation of an earthquake, as it might be used in a semantic network of news stories.<sup>1</sup>

At the same time that frame-based representations were being explored, a popular alternative approach was to use (some subset of) first-order predicate logic (FOL). A complete first-order logic allows one to make general statements about anonymous elements, with the introduction of variables as placeholders. In other words, in FOL it is possible to express general rules of inference that can be applied to all entities whose attributes satisfy the left-hand side of the  $\rightarrow$  inference operator. Thus, given the statement:

$$\forall x: \text{Man}(x) \rightarrow \text{Mortal}(x)$$

simply asserting *Man*(Socrates) entails *Mortal*(Socrates).

Since logic programming is based on the tenets of classical logic and comes equipped with automated theorem proving mechanisms, this approach allowed the development of inference systems whose soundness and completeness could be rigorously demonstrated. But while many of these early inference systems were logically sound and complete, they were often not very useful, as they could only be applied to highly proscribed areas or “toy problems.” The problem was that a complete first-order predicate logic is itself computationally intractable, as certain statements may prove *undecidable*.

In contrast, the frame representations offered a rich, intuitive means of expressing domain knowledge, yet they lacked the inference mechanisms and rigor that predicate logic systems could provide. Early efforts to apply predicate logics to frame representations soon revealed that the problem was computationally intractable. This occurred for two reasons: (1) the frame representation was too permissive; more rigorous definitions were required to make the representation computational; and (2) complete first-order predicate logic itself is intractable.

Several subsets of complete FOL have since been defined and successfully applied to develop useful computational models capable of significant reasoning. For example, the Prolog programming language is based on a subset of FOL that severely limits the use of negation. The family of description logic (DL) systems is a more recent development, and one that is especially well-suited to the development of ontologies, taxonomies, and controlled vocabularies, as an important function of a DL is as an auto-classifier.

Description logic can be viewed as a combination of the frame-based approach with FOL. In the process, both models had to be scaled back to achieve an effective solution. Like frames, the DL representation allows for concepts and relationships among concepts, including simple taxonomic relations as well as other meaningful types of association. Certain restrictions however, are placed on these relations. In particular, any relation that involves class membership, such as the *isa* or *inverse-isa* relations, must be strictly acyclic.

The predicate logic used in a description logic system is also limited in various ways, depending on the implementation. For example, the most minimal form of a DL does not allow any form of existential quantification. This limitation allows for a very easily computed solution

---

<sup>1</sup> This example is excerpted from *Artificial Intelligence*, by Patrick Winston, Addison-Wesley, 1984.



space but the resulting expressivity is severely diminished. The next step up in representational power allows limited existential quantification but without atomic negation.

Indeed, there is today a large family of description logics that have been realized, with varying levels of expressivity and resulting computational complexities. In general, DLs are decidable subsets of FOL, and the decidability is due in large part to their acyclicity. The theory behind these models is beyond the scope of this discussion; the interested reader is referred to *The Description Logic Handbook*, by Franz Baader, et al (Eds.), Cambridge University Press, 1993, ISBN number 0-521-78176-0.

The two main ingredients of a DL representation are *concepts* and *roles*. A major distinction between description logic and other subsets of FOL is in its emphasis on set notations. Thus, a DL concept never corresponds to a particular entity but rather to a *set* of entities, and the notations used for logical conjunction and disjunction are set intersection and union.

DL concepts can also be thought of as unary predicates in FOL. Thus, the expression: *Person*  $\cap$  *Young* can be interpreted as the set of all children, with the corresponding FOL expression:

$$Person(x) \wedge Young(x)$$

Syntactically then, DL expressions are variable-free, with the understanding that the concepts always reference sets of elements.

A DL role is used to indicate a relationship between the two sets of elements referenced by a pair of concepts. In general, DL notations are rather terse, and the concept (or set of elements) of interest is not explicitly represented. Thus, to represent the set of individuals whose children are all female, we would use:  $\forall$  *hasChild.Female*. The equivalent expression in FOL might be something like:

$$\forall x: hasChild(y, x) \rightarrow female(x).$$

In terms of set theory, a role potentially defines the Cartesian product of the two sets. Roles can have restrictions, however, that place limitations on the possible relations. A *value* restriction limits the type of elements that can participate in the relation; a *number* restriction limits the number of such relations an element can participate in.

In addition, each role defines a *directed* relation. For example, if *x* is the child of *y*, *y* is not also the child of *x*. In the above example *hasChild*, the parent concept is considered the *domain* of the relation, and the child is considered the *range*. Elements belonging to the set of objects defined by the range concept are also called role fillers. Number restrictions apply to the number of role fillers that are required or allowed in a relation. For example, a parent can be defined as a person having at least one child:

$$Person \cap (\geq 1 \text{ hasChild}).$$

A DL representation is constructed from a ground set of *atomic concepts* and *atomic roles*, which are simply asserted. *Defined concepts* and *defined roles* are then derived from these atomic elements, using the set operations of intersection, union, negation, etc. Most DLs also allow existential and universal quantifiers, as in the above examples. Note, however, that these quantifiers always apply to the role fillers only.

The fundamental inference operation in DL is *subsumption*, which is usually indicated with subset notation. Concept A is said to subsume B, or  $A \subseteq B$ , when all members of concept B are

contained in the set of elements defined by concept A, but not vice versa. That is, if B is a proper subset of A, then A subsumes B. This capability has far-reaching repercussions for vocabulary and ontology developers, as it enables the system to automatically classify newly introduced concepts. Moreover, correct subsumption inferencing can be highly nontrivial as, in general, this requires examining all of the relationships defined in the system and the concepts that participate in those relations.

### 2.3.1 The NCI Thesaurus Description Logic

The NCI Thesaurus is currently developed using the proprietary Apelon Inc. Ontylog™ implementation of description logic. Ontylog is distributed as a suite of tools for terminology development, management, and publishing. Although the underlying inference engine of Ontylog is not exposed, the implementation has the characteristics of what is called an AL<sup>-</sup> (attributive language) or FL<sup>-</sup> (frame language) description logic. It does not support atomic negation but does appear to provide all other basic description logic functionality.

The NCI Thesaurus is edited and maintained in the Terminology Development Environment (TDE) provided by Apelon. The TDE is an XML-based system that implements the DL model of description logic based on Apelon's Ontylog Data Model. The Data Model uses four basic components: *Concepts*, *Kinds*, *Properties*, and *Roles*.

As in other DL systems, concepts correspond to nodes in an acyclic graph, and roles correspond to directed edges defining relations between concept members. Each concept has a unique kind. Formally, *Kinds* are disjoint sets of concepts and represent major subdivisions in the NCI Thesaurus.

More concretely, *Kinds* are used in the role definitions to constrain the domain and range values of the triple. For example, a role *geneEncodes* might have its domain restricted to the *Gene\_Kind* and its range to the *Protein\_Kind*. “Findings and Disorders” and “Anatomy” are two additional examples of kinds.

As in all DLs, all roles are passed from parent to child in the inheritance hierarchy. For example, a “Malignant Breast Neoplasm” has the role *located-in*, connecting it to the concept “Breast.” Thus, since the concept “Breast Ductal Carcinoma” *is-a* “Malignant Breast Neoplasm,” it inherits the *located\_in* relation to the “Breast” concept. These lateral nonhierarchical relations among concepts are referred to as associative or semantic roles—in contrast to the hierarchical relations that reflect the *is-a* roles.

In the first-order algebra upon which Ontylog DL is based, every defined relationship also has a defined inverse relation. For example, if *A* is contained by *B*, then *B* contains *A*. Inverse relationships are useful and are expected by human users of ontologies. However, they have a computational cost. If the edges connecting concept nodes are bidirectional, then the computation becomes NP hard. Therefore, in the Ontylog implementation of DL, inverse relationships are not stored explicitly but are computed on demand.

## 2.4 Concept Edit History in the NCI Thesaurus

One of the primary uses of the NCI Thesaurus is as a resource for defining tags or retrieval keys for the curation of information artifacts in various NCI repositories. Since these tags are defined at a fixed point in time, however, they necessarily reflect the content and structure of the Thesaurus at that time only. Given the rapidly evolving terminologies associated with cancer research, there is no guarantee that a tag used at the time of curation in the repository will still

have the same definition in subsequent releases of the Thesaurus. In most cases the deprecation or redefinition of a previously defined tag is not disastrous, but it may compromise the completeness of the information that can be retrieved.

In order to address this issue, the EVS team has developed a *history* mechanism for tracing the evolution of concepts as they are created, merged, modified, split, or retired. (In the NCI Thesaurus, no concept is ever deleted.) The basic idea is that each time an edit action is performed on a concept, a record is added to a history table. This record contains information about relations that held for that concept at the time of the action as well as other information such as version number and timestamp, that can be used to reconstruct the state when the action was taken. Table 2.4-1 summarizes the information stored in the history table.

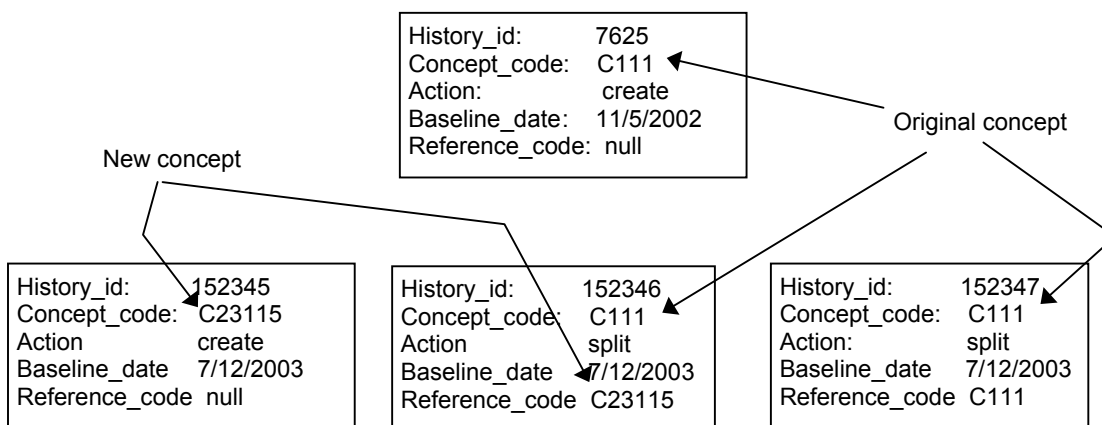
**Table 2.4-1 The NCI Thesaurus Concept History Table**

Column Name	Description
History_ID	A unique consecutive number for use as the database primary key
Concept_Code	The concept code for the concept currently being edited
Action	The edit action: {Create, Modify, Split, Merge, Retire}
Baseline_Date	The date of the NCI Thesaurus Baseline (see discussion below)
Reference_Code	The concept code for a second concept impacted by the action

The Reference\_Code column captures critical information concerning the impact of the edit actions on other concepts. This field contains the concept code of a second concept either participating in or affected by the editor's action. The value will always be null if the action is Create or Modify.

Capturing the history data for a Split, Merge, or Retire action is more complicated. In a Split, a concept is redefined by partitioning its defining attributes between two concepts, one of which retains the original concept's code and another that is newly created. This action is taken when ambiguities in the original concept's meaning require clarification by narrowing its definition.

In the case of a Split, three history records will be created: one for the newly created concept, (with a null Reference\_Code), and two for the original concept that is being split. In the first of these two records, the Reference\_Code is the code for the new concept; in the second one it is the code of the split concept.



**Figure 2.4-1 History Records for the Split Action**

For Merge actions, the situation is similar to a Split. In this case, two ambiguous concepts must be combined, and only one of the original concepts is retained. Again, there will be three history records created: two for the concept that will be retired during the merge, and one for the “winning” concept. The Reference\_Code in the history record for the “winning” concept will be the same as the Concept\_Code; i.e., the concept points to itself as a descendant in the Merge action. The Reference\_Code will be null in one of the entries for the retiring concept, while the second entry will have the code of the “winning” concept. Thus, this Reference code column points to the concept into which the concept in the Concept\_Code column is being merged.

Finally, if the action is Retire, there will be as many history entries as the concept has parent concepts. The Reference code column in these entries will contain the concept code of the parent concepts, one parent concept per history entry. The motivation for this is that end-users with documents coded by such retired concepts may find a suitable replacement among the concept’s parents at the time of retirement.

As the new EVS APIs support concept history queries, a minimal history will be added to all the vocabularies served from the Distributed Terminology Server (DTS) that are not edited by the EVS group. Concepts in vocabularies that are not edited by EVS will have a single history entry associated with them, a ‘create’ action with the date “May 1, 2003.” In the case of the NCI Thesaurus, concept history tracking has been ongoing internally since December 2002. However, for the purpose of publication in the DTS, a specific baseline has been selected to serve as “time zero” for concept history. This baseline is (internal) version 03.08c, which immediately preceded the NCI Thesaurus Version 2.0 released in caCORE 2.0. All of the concepts in this baseline have a ‘create’ action associated with them, dated “August 12, 2003,” the date of the 03.08c build.

## 2.5 The NCI Ontology Browser

All of the EVS vocabulary services are implemented as Oracle 8i relational databases accessed through [Apelon](#) server software. The NCI Thesaurus is a semantic model of the cancer namespace, built using description logic and served using Apelon’s DTS server software.

**NATIONAL CANCER INSTITUTE**

**Center for Bioinformatics**  
Welcome to the NCI DTS Browser

NCI DTS Browser is a modified version the servlet provided with Apelon's Distributed Terminology System (DTS). The Apelon DTS is a suite of run-time, middleware components that provides terminology services in a distributed application environment. The NCI DTS Browser can be used to search several namespaces. The NCI Thesaurus is a product of NCI. It is a description logic namespace that contains controlled terminology used at NCI. Using the NCI DTS Browser to search the terminology is easy, efficient and fast.

Please select a dictionary

<input checked="" type="radio"/> NCI Thesaurus	Published by NCI, this knowledgebase contains the working vocabulary used in NCI data systems. It covers clinical, translational and basic research as well as administrative terminology.	03.08c (08.12.03)
--	--	----------------------

Other Vocabularies:

<input type="radio"/> VA NDF	Published by the US Veterans' Administration, National Drug File covers clinical drugs used at the VA.	09.09.02
<input type="radio"/> LOINC	Published by the The Regenstrief Institute, the Logical Observation Identifier Names and Codes covers clinical laboratory terminology.	local 3MN (11.01.01)
<input type="radio"/> UWD Visual Anatomist	Published by the University of Washington, this knowledgebase covers human anatomy.	local (7.30.01)
<input type="radio"/> CTRM	Published by NCI, this knowledgebase currently contains human and mouse terminology. CTRM is an experimental knowledgebase in which ideas about representing vocabulary are tried out before possible inclusion in NCI Thesaurus.	11.01.02

**Connect**

Figure 2.5-1 The NCI Ontology Browser

Figure 2.5-1 shows the Welcome Page for the [NCI Ontology Browser](#). As indicated, the NCI Ontology Browser provides access to several other vocabularies in addition to the NCI Thesaurus. This section focuses specifically on the NCI Thesaurus, but the descriptions apply to the other vocabularies as well.

The first step in using the NCI Ontology Browser is to select the vocabulary to be accessed and to hit the **Connect** button. Depending on your internet settings, you may be redirected to a “Disabled Cookies Notification” page instructing you to enable “cookies.”<sup>2</sup> The NCI Ontology Browser uses session cookies to maintain and enhance the services provided.

To enable cookies on Internet Explorer, select **Tools → Internet Options** in the main menubar. When the pop-up dialog box appears, select the folder tab labeled *Privacy* to see the display shown in Figure 2.5-2(a). To reset cookie handling for just this site, select the **Edit...** button in the lower right-hand corner, and you will see something like the options shown in Figure 2.5-2(b).

Enter <http://nciterms.nci.nih.gov/NCIBrowser/> in the web site address box as shown in Figure 2.5-2(b) and click on **Allow**. The domain name *.nih.gov* should appear in the list of managed web sites with the setting “always allow.” Click **OK**, and then click “[here](#)” on the notification page to start the session.<sup>3</sup>

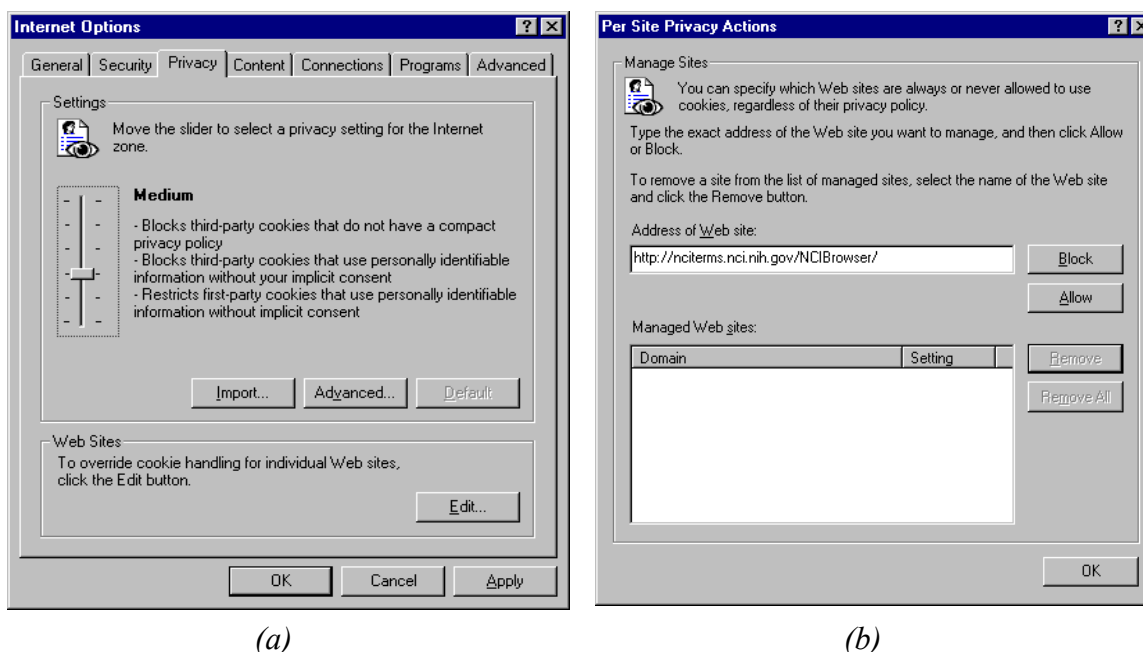


Figure 2.5-2 Resetting IE's Cookiehandling Options

<sup>2</sup> A cookie is a small piece of information that a web server requests your browser to store on your computer. This allows the server to recall specific information about your session while you are connected. There are two types of cookies, *permanent cookies* and *session cookies*. Permanent cookies are intended to retain your preferences in between sessions and are stored as files on your hard disk; session cookies last only for the duration of the session and are kept in memory by the web browser only while the browser is open. The NCI Ontology Browser uses only session cookies.

<sup>3</sup> Netscape users can negotiate similar settings by clicking on **Edit → Preferences → Privacy & Security → Cookies → ManageStoredCookies → CookieSites**.

Figure 2.5-3 displays part of the first screen you will see upon selecting the NCI Thesaurus in the NCI Ontology browser. A list of root concepts representing the major categories in the NCI Thesaurus is shown in the left-hand pane, and a search panel consisting of textboxes runs across the top, just to the right of the NCI logo.

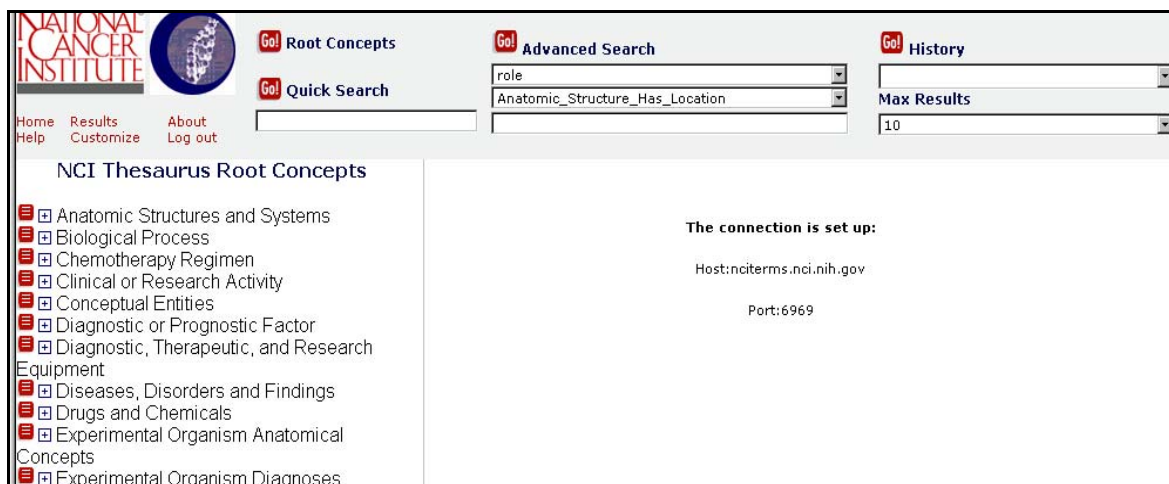


Figure 2.5-3 Browsing the NCI Thesaurus

The left-hand pane is also called the Taxonomy Viewer and is used for browsing the NCI Thesaurus concept tree. The pane to the right of the Taxonomy Viewer is used for displaying results and is referred to as the Display Panel. Finally, just below the NCI logo is a small menubar consisting of six options. Each of these components in the layout is described in more detail below.

### 2.5.1 The Taxonomy Viewer

The Taxonomy Viewer can be used to:

- traverse the vocabulary by clicking on the “+” (expand) and “–” (collapse) icons, and
- examine individual concepts in the Display Panel by clicking on the small red boxes with white lines running across them.

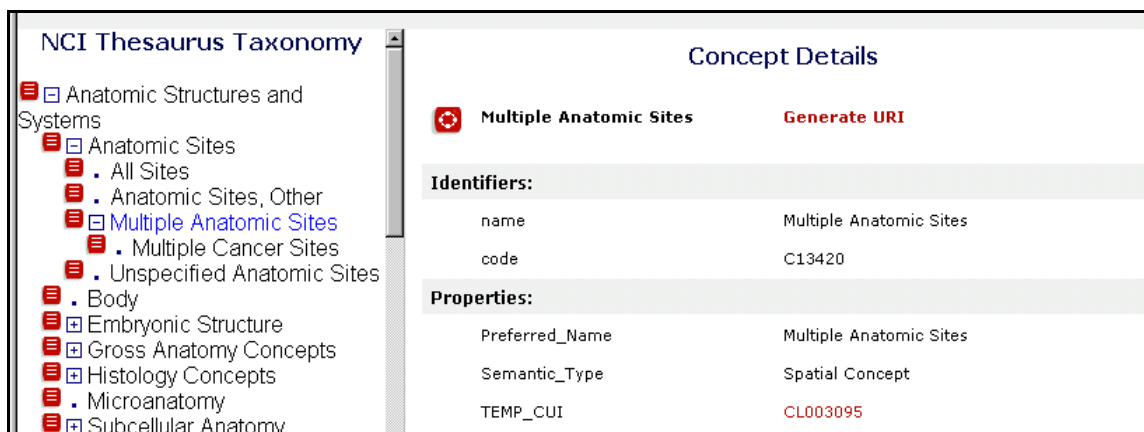


Figure 2.5-4 Selecting a Concept to Examine

Figure 2.5-4 shows a part of the display after fully expanding the **Anatomic Structures and Systems** node and then selecting **Multiple Anatomic Sites**. Note that the title of the Viewer panel no longer indicates that it is showing root concepts; the title now merely identifies the vocabulary, which in this case is the NCI Thesaurus. “Leaf” nodes—those having no subconcepts—have neither a + or – icon, as there are no descendants to expand. The concept currently detailed in the Display Panel is highlighted in blue in the Taxonomy Viewer.

## 2.5.2 The Search Panel

The search panel (Figure 2.5-5) provides five options for locating concept information.

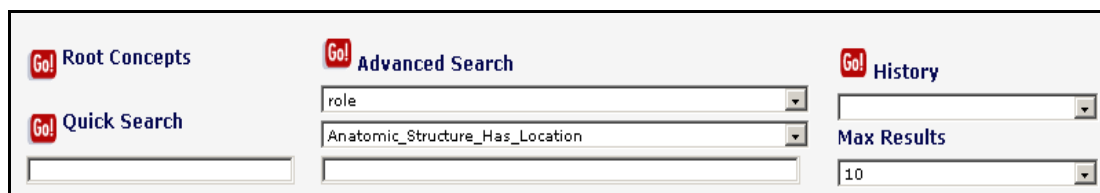


Figure 2.5-5 The Ontology Browser’s Search Panel

Each **Go!** button in Figure 2.5-5 initiates a different action. Clicking the **Go!** button next to Root Concepts will restore the Taxonomy Viewer to its original state, where only the (collapsed) root concept nodes are displayed. This option is provided for searches that are driven by selecting concepts in the Taxonomy Viewer, as described above.

Alternatively, if you know the name of the concept, you may wish to type this directly into the **Quick Search** textbox and click the **Go!** button just above that textbox. The Quick Search method is actually implemented as an advanced search—with all of the settable parameters for advanced search preset to default values. The Quick Search method results in a broad search that usually, but not always, will find all concepts matching the string you enter. The preset default values can be customized to enhance the Quick Search results, however, as described in the discussion of the menubar options.

The **Advanced Search** mode is accessed by the stack of widgets in the center of the search panel. Advanced Search allows you to further specify roles or properties that the concepts should exhibit. To perform an advanced search, you must first indicate which of these two attributes you wish to constrain by selecting from the first pull-down box.

As illustrated in Figure 2.5-6, when “role” is selected, the second pull-down box provides roles to select from, such as “Anatomic\_Structure\_Has\_Location.” By selecting a role, you are specifying that the concepts you are looking for should have that role. For instance, the concept should be an anatomic structure that *has* a location—not one that *is* a location.

Note that when using the Advanced Search interface to identify concepts that have a specific role (along with their accompanying target concepts), the DTS will show only the top-most concept in each *tree* of qualifying concepts. This was a design decision made by the software vendor to limit output, and assumes the user understands that all descendants of the concept being displayed also inherit both the specified role and the target concept.

Alternatively, when “property” is selected, the second box provides properties such as “CAS\_Registry.” By default, the value of the role or property to search for will be the first value

shown in this second pull-down box. To select a different role or property value, click the arrow on the lower pull-down box and select a new value from the list of recognized roles or properties.

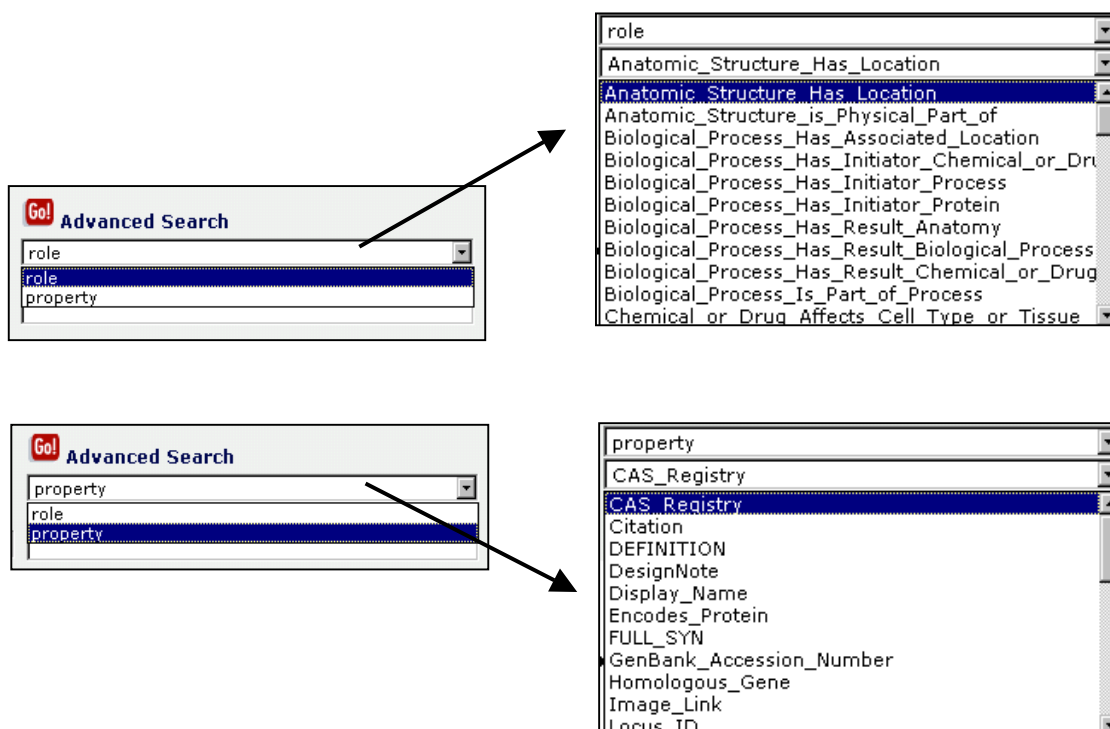



Figure 2.5-6 Advanced Search Options

After constraining the concept search by role or by property, you must now enter a text string in the textbox just below the two pull-down boxes. Clicking on the **Go!** button for Advanced Search then initializes the search, with the results displayed in the Display Panel. A more in-depth discussion of the advanced search mode is provided in the online NCI DTS Browser Guide's section on [Advanced Search Procedures](#).

The two remaining items on the search bar are the **History** and **Max Results** options. The History pull-down menu displays the concepts that have been viewed during the current session. Note, however, that examining a concept as one of a list of many results in the Display Panel does not constitute "viewing" that concept. In order to actually view a concept in a results list, you must select the details icon (  ) next to the concept. This will refresh the Display Panel with the Concept Detail Page for that concept.

In order to see recently visited concepts in the History list, you must click on the **Go!** button for history. Each time you click this button, any concepts that have been visited but are not yet listed there will be added to the pull-down list. Each concept in the History list can be revisited by selecting it from the pull-down list.

The Max Results option allows users to limit the number of results that will be returned, and is applied to both the Quick Search and Advanced Search methods.



### 2.5.3 The Display Panel

The Display Panel is used to show the results of both quick and advanced searches as well as to display detailed information about a selected concept. Figure 2.5-7 shows the contents of the Display Panel after executing a Quick Search.

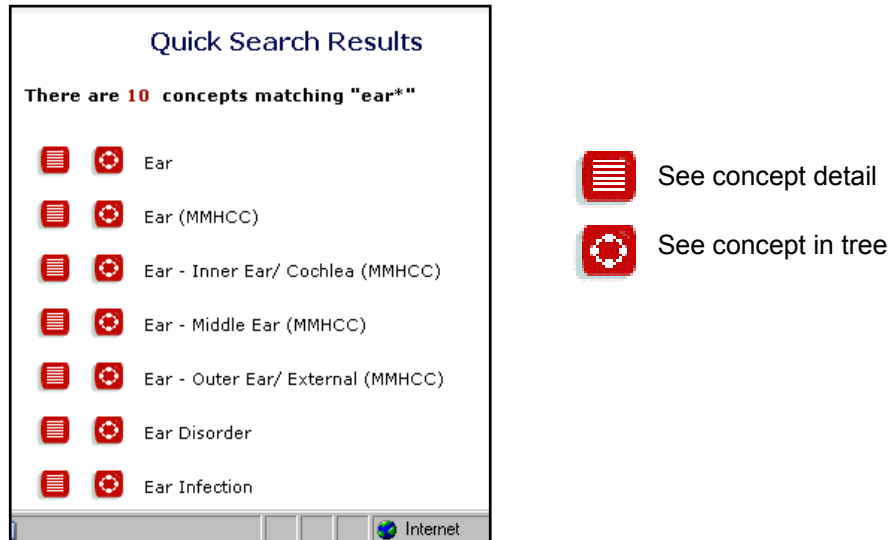


Figure 2.5-7 The Display Panel After a Quick Search

Each matching concept is listed by its concept name, with icons for (1) bringing up the Concept Details Page (Figure 2.5-8); or (2) exposing the concept node in the Taxonomy Viewer. The layout of the Advanced Search Results display is identical to that shown in Figure 2.5-7.



Figure 2.5-8 Concept Details Shown in the Display Panel

Clicking on the concept detail icon will refresh the Display Panel with detailed information for that concept. Figure 2.5-8 shows some of the Concept Details Page for the concept “ear.” The frame for the Taxonomy Viewer has been pulled to the far left, to show more of the Display Panel.

## 2.5.4 The Menubar

The menubar sits immediately below the NCI logo in the upper-left corner of the display and contains the six options described below.

- **Home** clears the Display Panel and restores the initial log in message.
- **Results** refreshes the Display Panel with the most recently displayed search results.
- **About** clears and displays version information for the DTS Browser in the Display Panel.
- **Help** opens an online version of the NCI DTS Browser Guide in a new window.
- **Customize** allows users to customize the default search parameters applied in Quick Searches; this option is discussed in more detail below.
- **Log Out** terminates the current session and returns you to the DTS Browser *Connection* window; the DTS Browser remains active.

Save

Customize

☐ Exact Match

Identifiers

☒ name
☐ code

Roles

Select All

Clear All

☐ Anatomic\_Structure\_Has\_Location
☐ Anatomic\_Structure\_is\_Physical\_Part\_of
☐ Biological\_Process\_Has\_Associated\_Location
☐ Biological\_Process\_Has\_Initiator\_Chemical\_or\_Drug
☐ Biological\_Process\_Has\_Initiator\_Process
☐ Biological\_Process\_Has\_Initiator\_Protein
☐ Biological\_Process\_Has\_Result\_Anatomy
☐ Biological\_Process\_Has\_Result\_Biological\_Process
☐ Biological\_Process\_Has\_Result\_Chemical\_or\_Drug
☐ Biological\_Process\_Is\_Part\_of\_Process
☐ Chemical\_or\_Drug\_Affects\_Cell\_Type\_or\_Tissue
☐ Chemical\_or\_Drug\_FDA\_Approved\_for\_Disease
☐ Chemical\_or\_Drug\_Has\_Associated\_Pathology
☐ Chemical\_or\_Drug\_Has\_Biochemical\_Class\_or\_Structure
☐ Chemical\_or\_Drug\_Has\_Biochemical\_Function
☐ Chemical\_or\_Drug\_Has\_Receptor
☐ Chemical\_or\_Drug\_Has\_Source
☐ Chemical\_or\_Drug\_Has\_Target\_Anatomy
☐ Chemical\_or\_Drug\_Has\_Target\_Organism
☐ Chemical\_or\_Drug\_Has\_Target\_Protein
☐ Chemical\_or\_Drug\_Plays\_Role\_in\_Biological\_Process
☐ Chemical\_or\_Drug\_is\_Part\_of\_Chemical\_or\_Drug
☐ Conceptual\_Part\_Of
☐ Disease\_Has\_Associated\_Anatomy
☐ Disease\_Has\_Associated\_Cell\_Type
☐ Disease\_Has\_Modifier

Figure 2.5-9 Customization Options in the Display Panel

Pressing **Customize** fills the Display Panel with a list of customization options—some of which are shown in Figure 2.5-9. Step-by step instructions for customizing the DTS Browser are provided in the NCI DTS Browser Guide in the section on [Customization](#).

## 2.6 The EVS Data Sources

The NCI Metathesaurus includes most of the UMLS Metathesaurus<sup>4</sup>, along with vocabularies developed internally at NCI (see Table 2.7-1) and external vocabularies that NCI has licensed. A limited model of the NCI Thesaurus is also accessible via the Metaphrase browser, as the NCI Source. The NCI Metathesaurus is available through the web interface described in this section, as well as through a Java API, which is described in the caCORE 2.0 Technical Guide.

**Table 2.6-1 NCI Local Source Vocabularies Included in the Metathesaurus**

<u>Vocabulary</u>	<u>Content</u>	<u>Usage</u>
NCI Thesaurus	Codes, keywords, and special purpose terminology for internal use at NCI	Reference terminology for internal NCI applications
NCIPDQ	Expanded and re-organized PDQ	CancerLit indexing and clinical trials accrual
NCISEER	SEER terminology	Incidence reporting
CTEP	CTEP terminology	Clinical trials administration
MDBCAC	Topology and morphology	Cancer genome research
ELC2001	NCBI tissue taxonomy	Tissue classification for genetic data such as cDNA libraries
ICD03	Oncology classifications	Cancer genome research and incidence reporting
MedDRA	Regulatory reporting terminology	Adverse event reporting
MMHCC	Mouse Cancer Database terminology	Mouse Models of Human Cancer Consortium
CTRM	Core anatomy, diagnosis, and agent terminology	Translational research by NCICB applications

---

<sup>4</sup> Certain proprietary vocabularies are of necessity excluded.

### 3.0 THE CANCER DATA STANDARDS REPOSITORY

The NCI [Cancer Data Standards Repository](#) is part of a broad initiative to standardize the common data elements (CDEs) used in cancer research data capture and reporting. It supports data management workflow and adherence to ISO/IEC 11179 metadata standards, the basis for its architecture. Defined by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), the purpose of the 11179 standard is to regularize the terms and representations used in the capture and subsequent annotation of shared electronic data. This type of information constitutes *metadata*—often characterized as “data about data.”

The sharing of electronic data generally requires that the user community agree upon the types of representation to be used, the manner in which alternate representations can be indicated, the intended meaning of such representations, and the names or “handles” by which such elements can be accessed. The ISO/IEC 11179 standard provides a framework for establishing, maintaining, and disseminating this type of information, and the caDSR is one particular implementation of this standard.

This chapter begins with a brief review of the ISO/IEC 11179 standard and its realization as the caDSR, an Oracle database and suite of metadata tools at NCI. The caDSR is currently used by a number of clinical trials groups for the management of metadata. Several web-based tools provide access to the caDSR, including:

- The [CDE Browser](#). This tool supports browsing, searching, and exporting of common data elements in the repository, and is accessible to the public.
- The [CDE Curation Tool](#). This tool supports the creation and modification of various components in the repository and requires an account with curatorial privileges.
- The [caDSR Administration Tool](#). This is the main administrative interface to all of the caDSR features and components. It is intended for use by caDSR administrators only, and requires an account with administrative privileges.
- The [Case Report Form CDE Review Response \(CCRR\) Tool](#). This tool was designed for sites conducting case report form compliance reviews with [CTEP](#) (the NCI Cancer Therapy Evaluation Program). Access to the CCRR Tool is limited to users engaged with CTEP holding accounts designated for this purpose.

A description of each of the first three tools follows the review of the ISO/IEC 11179 standard. The CCRR tool is not designed for general use and is beyond the scope of this document. A general note is that the phrase “common data element” is often used to emphasize that these elements are *common* across multiple domains and enterprises. The expressions “common data element” and “data element” are, however, equivalent, and should be interpreted as such.

#### 3.1 Modeling Metadata: The ISO/IEC 11179 Standard

Regardless of the application domain, any particular data item must have associated with it a variable name or tag, a conceptualization of what the item signifies, a value, and an intended interpretation of that value. For example, an entry on a case report form may be intended to capture the patient’s place of birth, and the corresponding value may be tagged electronically as “Patient\_placeOfBirth.” But what is the intended concept? Is the data element designed to

capture the country, the city, or the specific hospital where the person was born? Assuming that the intended concept is country, how is the resulting value to be represented electronically? Possible representations might include the full name of the country, a standard two- or three-letter abbreviation, a standard country code, or perhaps a specific encoding unique to the application.

Metadata is “data about data,” and refers to just this type of intentional information that must be made explicit in order to ensure that electronically exchanged data can be correctly interpreted. The purpose of the ISO/IEC standard is to define a framework and protocols for how such metadata can be specified, consistently maintained, and shared across diverse domains. The caDSR conforms to this standard; while it was developed specifically for the support of clinical trials data, usage of the caDSR is not limited to clinical applications.

The ISO/IEC 11179 standard defines a fairly complex model, and even the notion of metadata itself can be somewhat abstruse as it is a rather abstract concept. To facilitate understanding the model, this discussion uses a divide-and-conquer approach, and defines two very general types of components:

1. Information components whose purpose is to represent content; and
2. Organizational and administrative components whose purpose is to manage the repository.

This partitioning is not intrinsic to the ISO/IEC 11179 and, indeed, some of the components do not neatly fit into the separate categories. Nevertheless, it provides a useful framework.

The fundamental information component in the ISO/IEC 11179 model is the *data element*, which constitutes a single unit of data considered indivisible in the context in which it is used. Another way of saying this is that a data element is the smallest unit of information that can be exchanged in a transaction between cooperating systems. More specifically, a data element is used to convey the *value* of a selected *property* of a defined *object*,<sup>5</sup> using a particular *representation* of that value.

A critical notion in the metadata model is that any concept represented by a data element must have an explicit definition that is independent of any particular representation. In order to achieve this in the model, the ISO/IEC 11179 standard specifies the following four components:

1. A *DataElementConcept* consists of an *object* and a selected *property* of that object;
2. The *ConceptualDomain* is the set of all intended meanings for the possible values of an associated *DataElementConcept*;
3. The *ValueDomain* is a set of accepted representations for these meanings; and
4. A *DataElement* is a combination of a selected *DataElementConcept* and a *ValueDomain*.

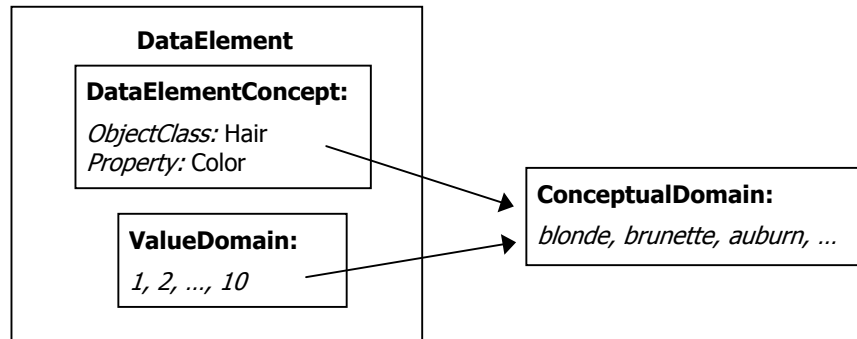
A simple example should help to clarify these definitions. Figure 3.1-1 shows a *DataElement* that might be used to represent hair color. The associated *DataElementConcept* uses the *ObjectClass* *Hair* and the *Property* *Color* to define the intended concept. The intended meanings for this data element are the familiar hair colors blonde, brunette, etc., but the *ValueDomain* uses a numeric representation that is mapped to these intended meanings. Both the *DataElementConcept* and the *ValueDomain* are components of the *DataElement*, and each

---

<sup>5</sup> The term *object* is used here in the sense defined by the ISO/IEC 11179 (see definition in Table 3.2-1) and does not have any literal correspondence to a caBIO Java object.

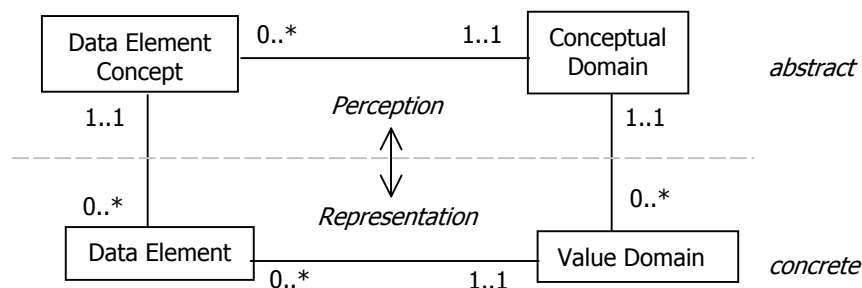
references the same `ConceptualDomain`, which is defined outside the `DataElement`. Important principles of this design are:

- The `DataElementConcept` is used to signify a concept *independent of representation*.
- The `ValueDomain` specifies a set of representational values *independent of meaning*.
- The `DataElement` combines a specific object and property with a value representation.
- The `ConceptualDomain` specifies the complete set of value meanings for the concept and allows the interpretation of the representation.



**Figure 3.1-1 Representing Data in the ISO/IEC 11179 Model**

Figure 3.1-2 uses a UML Class diagram to show the cardinality constraints that hold for these relations. Each `DataElement` must specify exactly one `DataElementConcept` and one `ValueDomain` in order to fully specify the data element. Similarly, each `DataElementConcept` and `ValueDomain` must specify exactly one `ConceptualDomain`. Conversely, a `ConceptualDomain` may be associated with any number of `ValueDomains` and any number of `DataElementConcepts`. Figure 3.1-3 shows an example of this, using the `color` property of different geometric objects as `DataElementConcepts`, and alternate color representations for the `ValueDomains`.

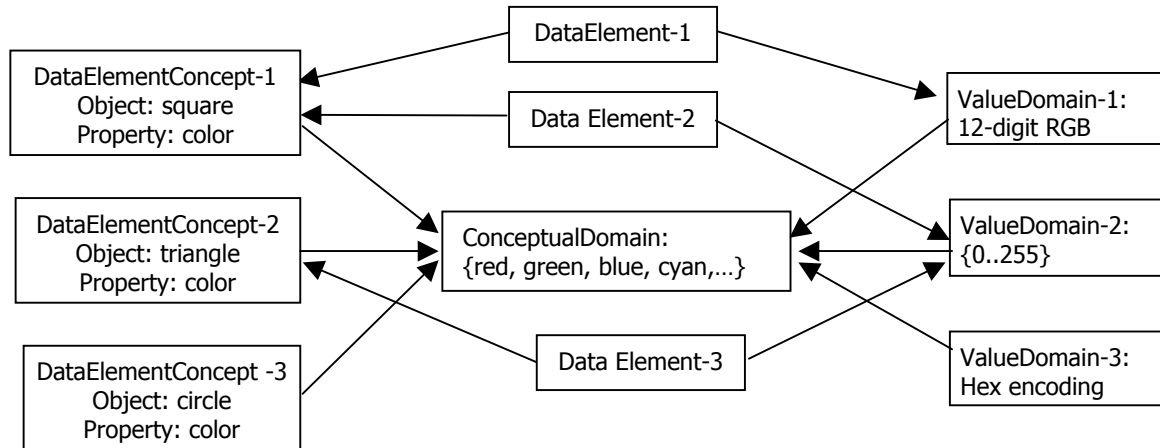


**Figure 3.1-2 Abstract and Concrete Components of the Data Representation**

A constraint not shown in any of these figures is that it is not possible to reuse the same `DataElementConcept`-`ValueDomain` pair to define a new `DataElement`, as this defines a logical redundancy. Thus, the “0..\*” cardinality constraints implied by Figure 3.1-2 are not quite as open-ended as they imply. Specifically,

- a `DataElement` specifies exactly one `DataElementConcept` and one `ValueDomain`;
- a `DataElementConcept` specifies exactly one `ConceptualDomain`;

- a ValueDomain specifies exactly one ConceptualDomain;
- a ConceptualDomain may be associated with any number of ValueDomains;
- a ConceptualDomain may be associated with any number of DataElementConcepts;
- a DataElementConcept may be associated with as many DataElements as there are ValueDomains (i.e., alternate representations) associated with the ConceptualDomain; and
- a ValueDomain may be associated with as many DataElements as there are DataElementConcepts associated with the ConceptualDomain.



**Figure 3.1-3 Many-To-One Mappings of Information Elements in the Metadata Model**

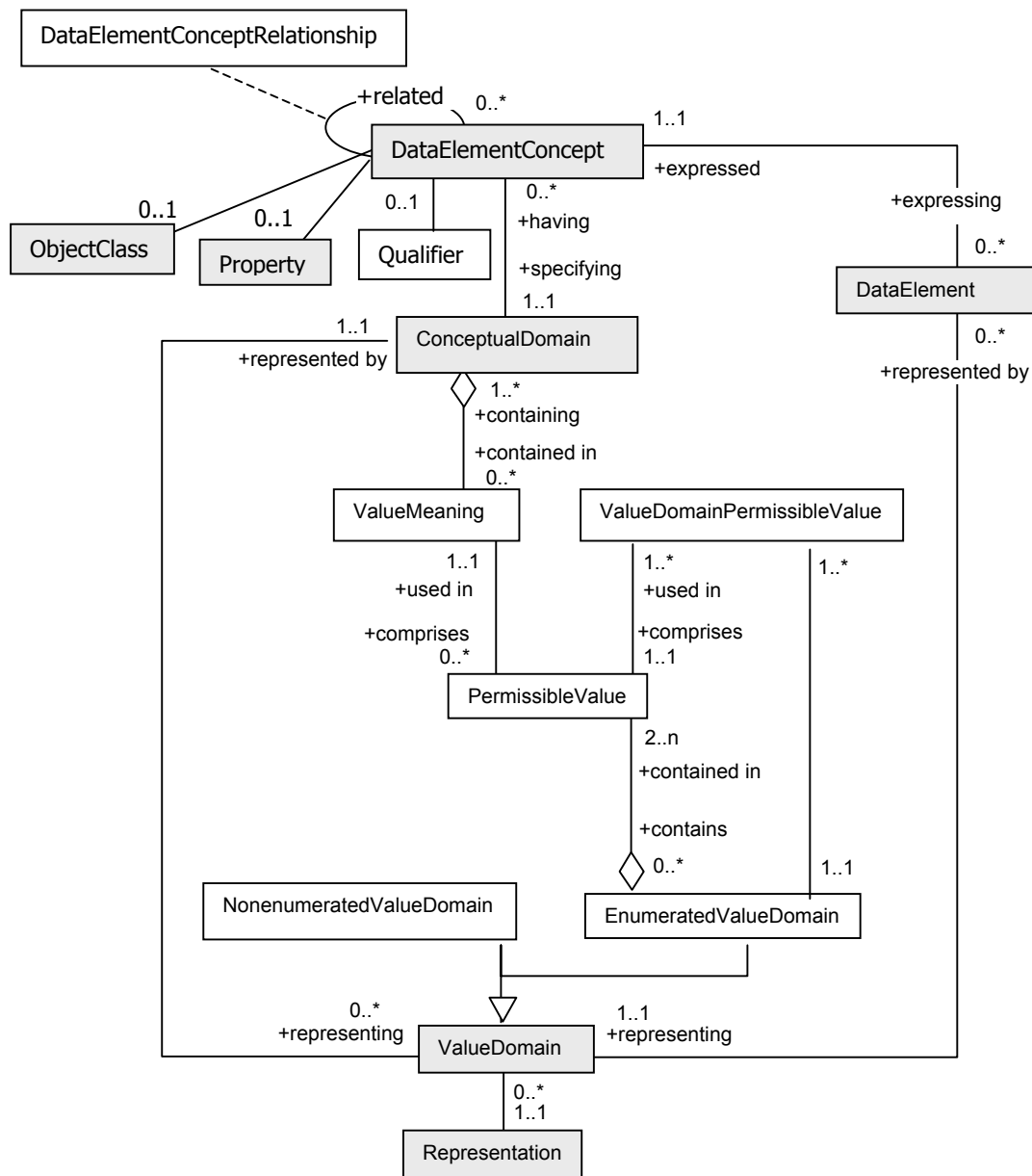
Many additional information components collaborate with these four core elements to provide the ISO/IEC 11179 infrastructure for content representation. These are described in the caDSR model in the next section, along with the organizational and administrative components that are used to document, classify, and, in general, manage the information components.

### 3.2 The caDSR Metamodel

Figure 3.2-1 again shows the four elements discussed thus far, but this time in the context of other components that collectively define the infrastructure for content representation. All of the components in Figure 3.2-1 that are highlighted in light gray must be *administered*. Pragmatically, this means that there is a formal protocol for creating these components, that there is an approval process in place for accepting newly proposed elements, and that there is a designated authority in charge of stewarding the component. Technically, this means that each of the highlighted components is an instance of an *AdministeredComponent*.

An *AdministeredComponent* is literally a component for which administrative information must be recorded. It may be a *DataElement* itself or one of its associated components that requires explicit specifications for reuse in or among enterprises—an *AdministeredComponent* is a generalization for all of the descendant components that are highlighted in Figure 3.2-1.

Several new components are introduced in Figure 3.2-1. Two of these are used to explicitly represent relationships between components defined in the ISO/IEC 11179 standard. A *DataElementConceptRelationship* is used to associate *DataElementConcepts* with one another; a *ValueDomainPermissibleValue* associates *PermissibleValues* with *EnumeratedValueDomains*. Table 3.2-1 provides concise definitions for all of the components in Figure 3.2.1.



**Figure 3.2-1 Information Component Infrastructure in the Metamodel**

Table 3.2-2 lists the attribute fields of an AdministeredComponent. The attributes listed there however, tell only half the story. Additional information about an AdministeredComponent derives from its associations with the organizational and administrative components depicted in Figure 3.2-2. Of these components, the only element that is also itself an AdministeredComponent is the *ClassificationScheme*.



**Table 3.2-1 Information Components in the caDSR Metamodel**

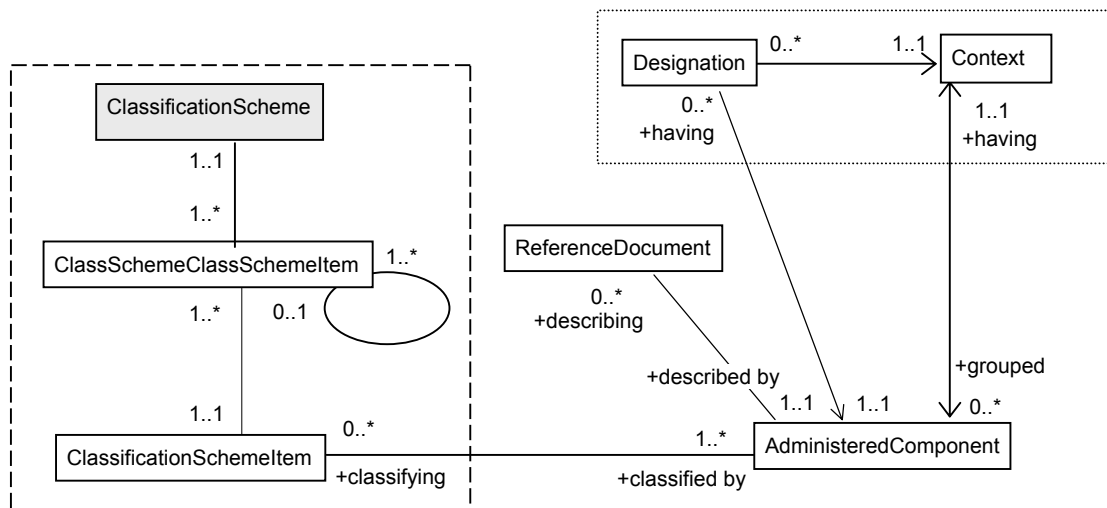
<b>Component Name</b>	<b>Definition</b>
<i>ConceptualDomain</i>	The set of all valid value meanings of a DataElementConcept expressed without representation.
<i>DataElement</i>	A unit of data for which the definition, identification, representation, and PermissibleValues are specified by means of a set of attributes.
<i>DataElementConcept</i>	A concept that can be represented in the form of a DataElement, independent of any particular representation.
<i>DataElementConceptRelationship</i>	An affiliation between two instances of DataElementConcepts.
<i>EnumeratedValueDomain</i>	A ValueDomain expressed as a list of all PermissibleValues.
<i>NonenumeratedValueDomain</i>	A ValueDomain expressed by a generative rule or formula; for example: "all even integers less than 100."
<i>ObjectClass</i>	A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behavior follow the same rules.
<i>PermissibleValue</i>	The exact names, codes, and text that can be stored in a data field in an information management system.
<i>Property</i>	A characteristic common to all members of an ObjectClass. It may be any feature naturally used to distinguish one individual object from another. It is conceptual and thus has no particular associated means of representation.
<i>Qualifier</i>	A term that helps define and render a concept unique. For example, given the ObjectClass <code>household</code> and the Property <code>annual income</code> , a Qualifier could be used to indicate <code>previous year</code> .
<i>Representation</i>	Mechanism by which the functional and/or presentational category of an item may be conveyed to a user. Examples: 2-digit country code, currency, YYYY-MM-DD, etc.
<i>ValueDomain</i>	A set of PermissibleValues for a DataElement.
<i>ValueDomainPermissibleValue</i>	The many-to-many relationship between ValueDomains and PermissibleValues values; allows one to associate a ValueDomain to a PermissibleValue.
<i>ValueMeaning</i>	The significance or intended meaning of a PermissibleValue.

**Table 3.2-2 Attributes of an *AdministeredComponent***

Attribute Name	Type	
id	String	required
preferredName	String	required
preferredDefinition	String	required
longName	String	optional
version	Float	required
workflowStatusName	String	required
workflowStatusDescription	String	optional
latestVersionIndicator	Boolean	required
beginDate	Date	optional
endDate	Date	optional
deletedIndicator	Boolean	optional
changeNote	String	optional
unresolvedIssue	String	optional
origin	String	optional
dateCreated	Date	required
dateModified	Date	required
registration	String	optional

Two “regions” are outlined in Figure 3.2-2: (1) the Naming and Identification region (upper right), and (2) the Classification region (lower left). The *ReferenceDocument* component is not included in either region. Each *AdministeredComponent* may be associated with one or more *ReferenceDocuments* that identify where and when the component was created and provide contact information for the component’s designated registration authority.

The purpose of the Naming and Identification region is to manage the various names by which components are referenced in different contexts. Many components may be referenced by different names depending on the discipline, locality, and technology in which they are used. In addition to the name attributes contained in the component itself (*preferredName*, *longName*), an administered component may have any number of alternative *Designations*. Each *Designation* is associated with exactly one *Context* reflecting its usage. The Classification region is used to manage classification schemes and the administered components that are in those classification schemes. Classification is a very fundamental and powerful way of organizing information to make the contents more accessible. Abstractly, a *ClassificationScheme* is any set of organizing principles or dimensions along which data can be organized. In the ISO/IEC 11179 model, a *ClassificationScheme* may be something as simple as a collection of keywords or as complex as an ontology. The *ClassificationScheme* element in Figure 3.2-2 is highlighted in light gray to reflect that it is an *AdministeredComponent*.



**Figure 3.2-2 Administrative and Organizational Components of the caDSR Metamodel**

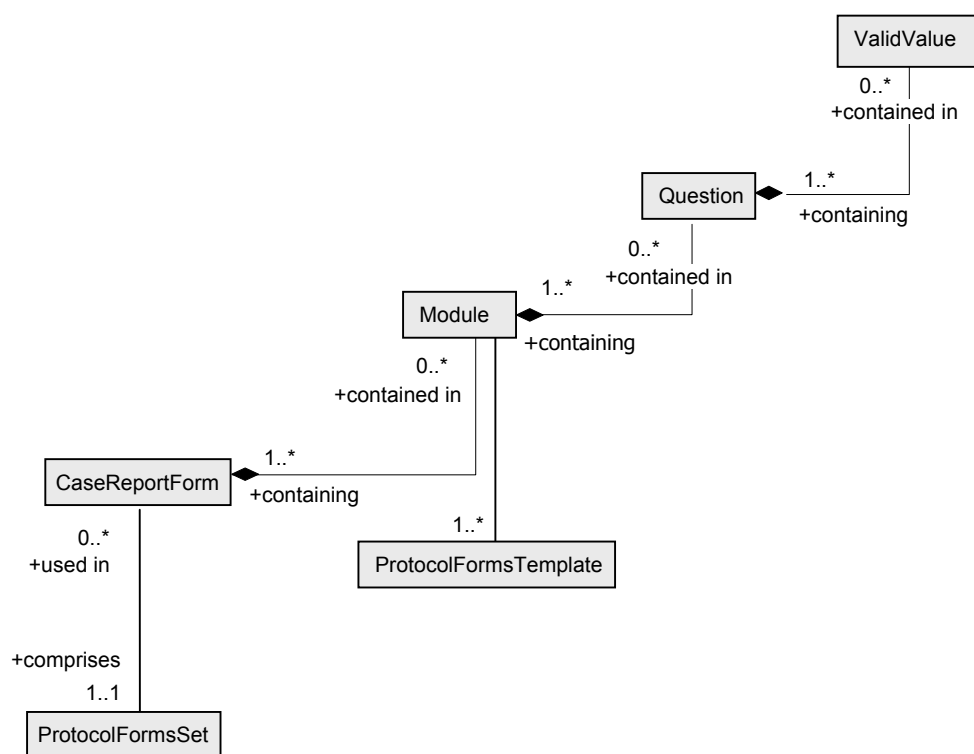
ClassificationSchemes that define associations among components can greatly assist navigation through a large network of elements; the associations may describe simple subsumption hierarchies or more complex relations such as causal or temporal relations. In particular, ClassificationSchemes with inheritance can enhance self-contained definitions by contributing the definition of one or more ancestors. In caDSR 2.0, only Data Elements and Data ElementConcepts may be assigned ClassificationSchemes; future plans include the ability to classify ConceptualDomains and ValueDomains.

The ClassificationScheme component serves as a container-like element that collects the classification scheme items participating in the scheme. In addition, the ClassificationScheme component identifies the source of the classification system and contains an indicator specifying that the scheme is alphanumeric, character, or numeric.

A *ClassificationSchemeItem* may be a node in a taxonomy, a term in a thesaurus, a keyword in a collection, or a concept in an ontology—in all cases, it is an element that is used to *classify* AdministeredComponents. It is quite natural for an AdministeredComponent that is used in different Contexts to participate in several ClassificationSchemes. ClassificationSchemes may coexist and a classified component may have a different name in each one, since each scheme is from a different context.

The *ClassSchemeClassSchemeItem* in the caDSR model is not a component of the ISO/IEC 11179 metamodel, but serves an important role in the implementation of the many-to-many mappings between ClassificationSchemeItems and ClassificationSchemes. This component is used to associate a set of ClassificationSchemeItems with a particular ClassificationScheme, and to store details of that association, such as the display order of the items within that scheme.

In addition to the caDSR components corresponding to elements of the ISO/IEC 11179 metamodel, the caDSR model defines a collection of domain-specific elements for capturing clinical trials data. All of the components described up to this point provide the infrastructure for managing shared data. The clinical trials components exercise the representational power of the metamodel, and are used to specify how clinical trials data should be captured and exchanged.



**Figure 3.2-3 Components in the caDSR Metamodel for Clinical Trials Data**

All of the components in Figure 3.2-3 are highlighted in light gray, as they are AdministeredComponents designed for use in NCI-sponsored clinical trials. Note that because these elements are *not* part of the ISO-11179 specification, they are not, technically speaking, ISO administered components. This caDSR design decision was made to ensure that the shared DataElements could be stewarded and controlled adequately. Definitions for these clinical trials metadata components are provided in Table 3.2-3.

NCI-sponsored clinical trials programs can populate the registry with instances of these components as needed to specify the metadata descriptors needed for that program. Programs currently participating in this effort include:

- The Cancer Therapy Evaluation Project ([CTEP](#))
- Specialized Programs of Research Excellence ([SPORes](#))
- The Early Detection Research Network ([EDRN](#))
- The Division of Cancer Prevention ([DCP](#))
- The Cancer Imaging Program ([CIP](#))
- The Division of Cancer Epidemiology and Genetics ([DECG](#))
- The Cancer Bioinformatics Infrastructure Objects Project ([caBIO](#))



**Table 3.2-3 Components in the caDSR Metamodel For Clinical Trials Data**

<b>Component Name</b>	<b>Component Description</b>
<i>CaseReportForm</i> (CRF)	A questionnaire that is a collection of data elements used to document patient information stipulated in the protocol. A CRF is used by clinicians to record information about patients' visits in a clinical trial.
<i>Question</i>	The text that accompanies a data element on a CRF; used to clarify the information being requested.
<i>Module</i>	A logical grouping of data elements on a CRF.
<i>ProtocolFormsSet</i>	A specific clinical trial protocol document and its collection of associated CRFs. Clinical trial protocols, along with their associated CRFs, stipulate the execution of clinical trials. A protocol is uniquely identified by a protocol ID, protocol version, and Context name.
<i>ProtocolFormsTemplate</i>	A boilerplate collection of components (modules, questions and valid values) to be included in a Case Report Form. The template form is not associated with any particular clinical trial.
<i>ValidValue</i>	An allowable value for a data element (question) on a CRF. <sup>6</sup>

### 3.3 The CDE Browser

The [CDE Browser](#) provides an excellent starting point for exploring the contents of the caDSR. This tool supports browsing, searching, and exporting of CDEs, and no explicit user account is required to use the tool. Figure 3.3-1 shows the starting page that is generated when you first enter the CDE Browser.

As shown there, the leftmost panel provides a navigable tree of contexts. As described in the Section 3.1, one of the fields in the ISO/IEC 11179 model is called *Context*. This field is used in the caDSR to distinguish CDE development efforts that are managed by different authorities. Each [Context](#) has a curatorial authority that manages the creation, editing, and designation of CDEs for that Context.<sup>7</sup>

Two icons are used to signify that the corresponding node either contains subnodes or is a final “leaf node” of the tree. A yellow folder icon () with a “+” sign in the lower right-hand corner indicates that the node can be expanded by clicking on the icon. A yellow folder icon with a “–” sign in the lower right-hand corner indicates that the node has been fully expanded and all of its subnodes are exposed. In this case, the node can be collapsed by clicking on the folder icon. A file icon () indicates that the node is a leaf and cannot be expanded; clicking on this icon has no effect.

<sup>6</sup> Although initially distinct, a ValidValue is now equivalent to a PermissibleValue

<sup>7</sup> The Context with the largest body of CDEs that have been approved for use in NCI clinical trials is managed by the NCI Cancer Therapy Evaluation Program. Additional discussion of the ISO/IEC 11179 standard, along with a specification of the business rules and conventions implemented by CTEP in developing CDEs, can be found in a downloadable document named [CTEP\\_CDE\\_Intro.doc](#) at the caDSR home page.

Contexts are used to limit the search for CDEs to those defined within that context only. A context is selected by clicking on the brown hypertext identifying the source for that context. The path to the currently selected context is always displayed in the top of the right-hand search pane.



Figure 3.3-1 The CDE Browser Welcome Page

The right-hand side of the interface holds the Data Element Search pane. This pane allows you to further restrict your data element search by entering additional constraints. Clicking on a context *before* entering any additional constraints in this search pane will immediately trigger a search for *all* CDEs defined in that context. As some of the contexts (such as the CTEP context) are fairly large, it is a good idea to enter constraints in the search pane before selecting a context.

In addition to its pull-down menus and textboxes, the search pane provides two buttons. The **Search Data Elements** button initiates the search. Selecting this option *without* specifying a context will find all CDEs satisfying the search pane constraints in *all* contexts. Similarly, selecting this option without first specifying either a context *or* any constraints in the search pane will find all CDEs in the caDSR. The **Clear** button will clear any selections in the search pane, but will preserve the current result set as well as the currently selected context.

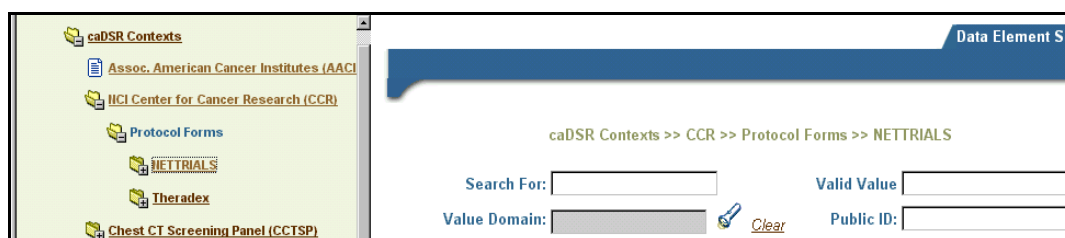


Figure 3.3-2 Display of the Currently Selected Context in the Search Pane

Contexts can be switched at any time; selecting a new context effectively de-selects the current context. In general, it is not possible to select multiple contexts unless they are subnodes of the same parent in the context tree. In this case, selecting the parent will simultaneously select

all contexts defined by the subnodes. Figure 3.3-2 shows a portion of the screen after making the selection “caDSR Contexts → CCR → Protocol Forms → NETTRIALS” in the context tree.

The basic principles of using the CDE Browser can be summarized as:

1. In the context tree, click on closed folder icons (“+”) to expand those nodes and explore data element groupings. Click on open folder icons (“-”) to collapse those branches in the tree.
2. The Data Element Search pane provides pull-down menus and textboxes for entering additional constraints to narrow the search for elements within the selected context. Execute the search by clicking on the **Search** button or selecting a context in the tree.
3. Clicking on a brown hypertext label in the context tree will select that grouping of data elements as the effective search domain, and will trigger the retrieval of data elements in that group that satisfy any constraints defined in the search pane.
4. The result set can be discarded or downloaded using either the Excel or XML formats by clicking on the appropriate download buttons. For users choosing to download XML formatted data, the document type definition (DTD) describing these documents can be downloaded from <http://ncicb.nci.nih.gov/NCICB/xml/dtds/cadsr>.

The options in the Data Element Search pane—including the pull-down menus, textboxes, and download features—are discussed below. Before considering these issues, however, the next section discusses the underlying structure and interpretation of the context tree in more detail.

### 3.3.1 The Structure of the Context Tree

Only the topmost nodes in the context tree correspond to actual contexts in the caDSR. All nodes occurring beneath the first level are branches that correspond to logical groupings of the data elements within that context. Groupings can be any of the following types:

- **Classifications and Classification Scheme Items.** A Classification is a collection of related Classification Scheme Items that define a subhierarchy within the tree. Data Elements, in turn, can be assigned to one or more Classification Scheme Items. Classification assignments are managed by the Context administrator, and only those classification schemes whose workflow status is ‘RELEASED’ are displayed in the tree.
- **Protocol Forms.** Protocol Forms are the case report forms used in a clinical trial. Each such form is composed of data elements, and clicking on the corresponding node in the context tree will return the data elements for that form in the Data Element Search pane.
- **Protocol Form Templates.** Where possible, case report forms are generalized to produce reusable protocol form *templates*. In the CTEP Context, these templates are grouped by the phase of the clinical trial and by the disease studied. Within each of these two groups, they are further classified by the type of template. Clicking on one of these template form names in the tree will return the data elements for that template in the Data Element Search pane.
- **Specialized Groupings.** Other types of groupings can be found in the tree as well. For example, under each type of disease classification, the CTEP Context groups data elements into “core” and “non-core” categories.

### 3.3.2 Data Element Search Pane

Figure 3.3-3 shows the right portion of the CDE Browser page, the Data Element Search pane, which provides additional fields to further limit the search.

The screenshot shows the 'Data Element Search' pane with the following fields and options:


- Search For:** Text input field.
- Valid Value:** Text input field.
- Value Domain:** Dropdown menu with a search icon and a 'Clear' link.
- Public ID:** Text input field.
- Data Element Concept:** Dropdown menu with a search icon and a 'Clear' link.
- Classification:** Dropdown menu with a search icon and a 'Clear' link.
- Version:** Radio buttons for 'Latest Version' and 'All Versions'.
- Context Use:** Dropdown menu.
- Workflow Status:** Dropdown menu with options: ALL, APPRVD FOR TRIAL USE, CMTE APPROVED, CMTE SUBMTD.
- Search Field(s):** Dropdown menu with options: ALL, Long Name, Preferred Name, Document Text.
- Buttons:** 'Search Data Elements' and 'Clear'.
- Note:** 'Wildcard character for search is \*'.

Figure 3.3-3 The Data Element Search pane

- **Search For:** This text field is coupled to the **Search Field(s)** pull-down menu appearing in the lower right-hand corner: Keywords entered in the text field are matched only to those selected fields in the data elements. The defaults are to search all fields for an exact match to the entered keywords. These defaults can be overridden by selecting specific fields and/or using an asterisk (\*) to signify wildcard matching. Multiple search fields can be selected from the pull-down list by holding the **Ctrl** key.
- **Valid Value:** Each data element has a set of valid values associated with it, and this field provides a means of filtering data elements according to their valid values. As with the keyword search field, wildcard characters can be used to broaden the search. For data elements having an enumerated value domain, the valid values are included in the permissible values of that domain.
- **Value Domain:** The Value Domain of the data element indicates the permissible values that can be collected in an actual research study. Value Domains must be referred to by name, and as this selection criterion does not support free text, the Value Domain must be selected from a controlled list of names. Clicking on the 🔍 icon will generate a pop-up window that performs a keyword search for Value Domains (see Figure 3.3-4). Selecting a particular Value Domain from the results appearing in the pop-up window will insert that name as the Value Domain constraint in the Data Element Search pane.
- **Data Element Concept:** The Data Element Concept indicates the semantic meaning of the associated data element. To filter data elements based on Data Element Concept, click on the 🔍 icon and proceed as described above for Value Domains.
- **Workflow Status:** The Workflow Status specifies the administrative status of a data element. In most cases you will probably want to search for data elements with a status of RELEASED in the name, but any status is available as a search criterion. To filter data



elements based on their Workflow Status, use the pull-down menu to select the Workflow Statuses of interest. Multiple workflow statuses can be selected by holding down the **Ctrl** key while clicking.

- **Public ID:** To filter data elements based on Public ID, enter a valid Public ID in the field provided.<sup>8</sup> Wildcard characters (\*) can be used at the beginning, middle, or end of the search term to broaden the search.
- **Classification:** Classification Scheme Items are particular classifications that data elements have been assigned to by the Context administrator. To filter data elements based on their Classification Scheme Item assignments, click on the  icon to search for Classification Scheme Items, and proceed as described above in selecting Value Domains.
- **Version:** Data Elements are assigned version numbers by the Context administrator. To filter data elements based on version numbers, select the **Latest Version** or **All Versions** radio button.
- **Context Use:** A data element “belongs” to the Context that “owns” it, and the owning context has the sole administrative authority to edit and update that element. But in addition to belonging to an owning Context, a data element can also be “designated” for use by another *non-owning* Context. The default search behavior is to search for data elements that are *Owned & Used by* the currently selected context. Using the **Context Use** option however, it is also possible to search for data elements that are simply *Used by* or *Owned by* a particular context.
- **Search Field(s):** This option defines how the keywords entered in the **Search for** textbox should be used to identify matching data elements. The default is to match the keywords to all of the fields listed in the pull-down menu. This can be overridden by selecting one or more fields from the menu; multiple search fields can be selected from the pull-down list by holding the **Ctrl** key.

### Value Domains

Please enter a keyword. This search will display all value domains which have the search criteria in their long name or preferred name. Wildcard character is \*.

**Keyword**

**Restrict Search to Current Context** ☒

Preferred Name	Long Name	Context	Workflow Status	Definition
<a href="#">DISEASE_RELATN</a>	Disease related symptoms	CCR	DRAFT NEW	The Value Domain for Disease related symptoms
<a href="#">DISEAS_STG</a>	Stage of Disease	CCR	DRAFT NEW	The Value Domain for Stage of Disease
<a href="#">DZ_DETECT_METH</a>	Disease Detection Methods	CCR	DRAFT NEW	Methods and approaches used to detect and document sites of disease

**Figure 3.3-4 Constraining Data Elements by Value Domain**

<sup>8</sup> Starting with version 2.0, the CDE ID has been replaced with Public ID.

After setting all of the appropriate search criteria, the search can then be executed by clicking the [Search Data Elements](#) button, or alternatively, by selecting a new data element grouping from the context tree. In either case, the currently selected context as well as any specified criteria in the search pane will be applied to the result set. The results will then be listed in a table just below the download options.

### 3.3.3 Working with the Results Set

The results table is displayed with a maximum of 40 records per page. When the record count exceeds 40, the results are paginated, and the CDE Browser provides two ways to scroll through the entire set. First, simple arrow keys at both the top and bottom of each page provide a mechanism for advancing or regressing in single page increments. In addition, a pull-down menu sandwiched between the [previous](#) and [next](#) buttons provides a complete list of all of the result pages, thus allowing the user to skip ahead or back to any page whatsoever.

Clicking on the first field in the results table—*Preferred Name*—brings up a new window detailing the selected data element. The menu bar running across the top of this window contains selectable tabs that link to additional pages detailing the Data Element Concept, Valid Values, Classifications, and Usage for the selected data element. The first tab is linked to the Information Page for the data element itself and is, by default, the initial display. If a different data element is subsequently selected from the results table while this Details window is still open, the pages will be refreshed with information for that newly selected data element. Thus, only one data element Details window can be displayed at a time.

Several formats are provided for capturing result sets. Clicking on the [Download Data Elements to Excel](#) link saves the results in an Excel Comma-Separated Values (CSV) file, which can be opened as an Excel spreadsheet. After clicking on the link, a File Download window will be displayed with the options of opening the file immediately or saving it to disk. Selecting the [Open](#) button will start up Excel in a separate window, displaying the file you have just downloaded. The first few columns of the Excel table are identical to the CDE Browser display. The subsequent columns provide a condensed version of the information contained in the Details window.

Alternatively, the result set can be saved in XML format by clicking on the [Download Data Elements as XML](#) link. The XML file building process is considerably slower than that for Excel, and the notification window warning that the download may take a few minutes must be left open until the File Download window is displayed. Directly opening the XML output appears to crash some browsers, and the Save option is strongly recommended.

Searching for data elements that hold Protocol Forms Templates in the context tree produces additional results. As in the foregoing discussion, selection criteria in the search pane can be set to limit the search. The result set is again displayed in a multi-page results table, and the Details window can also be invoked by selecting a Preferred Name for a data element in that table. But in this case, a [Download Template](#) option is provided in addition to the usual Excel and XML download options. Clicking on this link will display the Word document rendition of the template in a separate window.

Online help covering much of the same material discussed here is also available on the CDE Browser pages.

### 3.4 The CDE Curation Tool

Like the CDE Browser, the CDE Curation Tool allows users to browse administered components in the registry. In addition, the Curation Tool allows designated users to create and/or modify existing components.<sup>9</sup> Table 3.4-1 summarizes the browsing and editing capabilities for the various components accessible through the CDE Curation Tool.

**Table 3.4-1 Component-Specific Browsing/Editing Capabilities in the CDE Curation Tool**

	Search	Create	Edit
Data Element	✓	✓	✓
Data Element Concept	✓	✓	✓
Value Domain	✓	✓	✓
Permissible Value	✓	✓	
Conceptual Domain	✓		
Classification Scheme Item	✓		
Data Elements matched to Case Report Form Question	✓		

The dark blue navigation bar running across the top of all screens for the Curation Tool provides buttons for switching between Search, Create, and Edit modes; for accessing online help; and for terminating the session (see Figure 3.4-1).



**Figure 3.4-1 The CDE Curation Tool's Navigation Bar**

Extensive context-sensitive help is available online for the Curation Tool. To access context-specific help, highlight the text field, pull-down list, button, or hyperlink you would like help on, and press the **F1** key. To access a hyperlinked table of contents to the help pages, click on the **Help** button on the navigation bar. Over 100 pages of step-by-step instructions are available online. The discussion that follows is a condensed version of the material presented there.

#### 3.4.1 Searching for Administered Components

The CDE Curation Tool's interface for searching is not case-sensitive and supports wild card matching via the asterisk (\*) character. For example, entering "a\*" in the search terms textbox will match all of the selected search fields whose text begins with either "A" or "a."

Like the caDSR Admin Tool, the CDE Curation Tool provides search access to the caDSR components as well as to terms and definitions in the NCI Thesaurus and the NCI Metathesaurus.

<sup>9</sup> While there is some redundancy among the caDSR tools, each provides slightly different capabilities. See Table 4.6-1 for a comparative summary.

This capability allows users to use controlled vocabulary terms and definitions in the creation of new Administered Components, thus streamlining the harmonization process for newly introduced nomenclatures and their interpretations. Access to these EVS vocabularies is available in the search/browse modes as well as in the edit/create modes of the Curation Tool.

Figure 3.4-2 shows the basic search window for Data Elements, with annotations from the online help documentation. As indicated there, the initial screen displays the default column headers that will appear in the results table after a search is executed. These defaults can be modified to include whatever information is needed and to exclude those columns that are not relevant, using the **Display Attributes** menu (option 5) in the left sidebar.

The left sidebar contains slots and widgets for specifying search criteria. Figure 3.4-2 shows the options that are displayed when the type of component being searched for is a Data Element.

- **Search For** (pull-down): Specifies the *type* of administered component to search for.
- **Search In** (pull-down): Specifies the fields to search on in those components.
- **Search Terms** (textbox): Specifies keywords to match these fields to.
- **Filter By** (pull-down): Constrains the Context and Workflow Status of the matching components. The context box permits only single selections; the status box allows multi-selections. To select multiple workflow statuses, hold down the control key (**Ctrl**) while making selections. Only those matching components that also satisfy the context and workflow status specified in these slots will be returned; the default behavior is to search all contexts and to include all workflow statuses.
- **Display Attributes** (pull-down): Controls the information that will be included in the results table. Additional columns can be added and any of the default columns can be removed.
- **Start Search** (button): Executes the search.

The screenshot shows the 'CDE Curation Tool' interface. The sidebar on the left contains the following sections:

- 1) Search For: A pull-down menu currently set to 'Data Element'.
- 2) Search In: A pull-down menu currently set to 'Names and Definition'.
- 3) Enter Search Term: A text input field with a placeholder 'use \* as wildcard'.
- 4) Filter By: Two pull-down menus. The first is 'Owned By and Used By' set to 'All Contexts'. The second is 'Workflow Status' with a multi-select list containing 'All Statuses', 'APPROVD FOR TRIAL', and 'CMTE APPROVED'.
- 5) Display Attributes: A pull-down menu showing 'Preferred Name', 'Owned By', 'Used By', and 'Version'.

The main area of the tool displays a table of search results. Above the table, a red box contains the text 'Search results are displayed in table format here'. The table has the following columns: 'Preferred Name', 'Owned By', 'Used By', 'Version', 'Long Name', 'Workflow Status', and 'Definition'. A red arrow points to the 'Start Search' button at the bottom of the sidebar with the text 'Click here to start your search'.

Figure 3.4-2 The CDE Curation Tool's Basic Search Window

The selectable values in the pull-down menus, and, in some cases, the options themselves, vary according to the type of component. By definition, the display attributes shown in option (5) reflect component-specific properties. A second option that is often component-dependent is the **Filter By** alternatives. For Data Elements, Data Element Concepts, Value Domains, and Conceptual Domains, the options are exactly as shown in Figure 3.4-2. For Classification Scheme Items and Permissible Values however, the “Workflow Status” feature of the **Filter By** option is replaced with “Classification Scheme” and “Conceptual Domain,” respectively.

The search interface for CRF Questions removes option 3 (**Search Terms**). This search function automatically retrieves any Case Report Form (CRF) questions marked with a Reviewer Action of “Draft New” in the CDE Compliance Review Tool (CRT). This search function is used in the CRF review process for CDE compliance, and returns only those questions marked as “Draft New” in the CRF under review.

Most of the search criteria options in the sidebar are self-explanatory. One option that merits further explanation is the **Search For** “Data Elements” (option 1), when option 2 is set to **Search In** “Protocol ID/ CRF Name.” This search function retrieves Data Elements used as matches to submitted Questions on CRFs during the CDE compliance review process.

If the goal is to find only those Data Elements used as matches in a *particular* case report form, the **Search Terms** field should identify that CRF name explicitly. Alternatively, entering a Protocol Id will find all Data Elements used as matches in *all* case report forms contained in that protocol. Finally, the CRF Name search allows wildcard (“\*”) matches, but the Protocol ID search requires an exact match.

### Working with the Search Results

The search results appear in table format on the right side of the Search window, under the column headings selected in the **Display Attributes** option. Clicking on a column heading in the table causes the rows to be re-sorted (alphabetically or numerically) on that column. It is also possible to add or remove columns *after* the search, by selecting or removing attributes using the sidebar’s display option. To have these changes take effect on the current result set, click the **Update** button.

Additional display and editing options are accessible in the results menubar that runs above the results table and just below the navigation bar. The options presented in the results menubar vary depending on the type of administered components contained in the result set. Each result row has a leftmost checkbox, and selecting any one or more rows will enable this menubar. Figure 3.4-3 shows a screen shot where the results menubar for Data Element components has been enabled by selecting the first row in the results table.

In this case, the first button is **Edit Selection**, and all of the remaining buttons are enabled. Selecting more than one row changes the edit button to **Edit Block** and disables the **Details** and **Get Associated** buttons. Further consideration of the edit capabilities available from this search screen is deferred to the more general discussion of Creating and Editing Administered Components.

The second button is labeled **Designations** and allows users to add *designations* to the selected component(s). Designating simply means that you will “point” the existing Data Element, Data Element Concept or Value Domain (including all of its attributes) to a new Context. In other words, the designated Administered Component will now be “Used By” that Context.



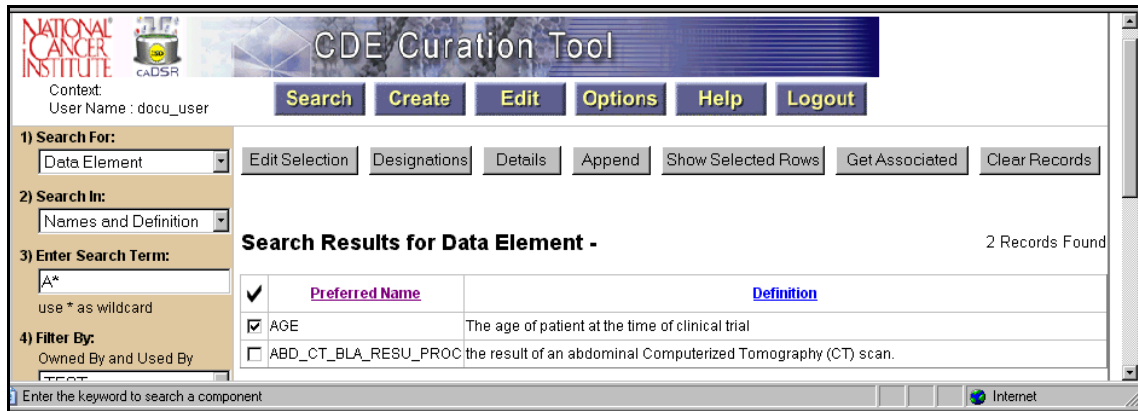


Figure 3.4-3 Enabling the Search Results Menubar

Only those components having the following workflow statuses can be designated for use in other contexts:

- Released
- Approved for Trial Use
- Committee Submitted Used
- Draft Modified
- Committee Approved
- Released non-compliant

Pressing the **Designations** button will cause a pop-up dialog box to appear (see Figure 3.4-4) with a pull-down menu for selecting the context to which you would like to add a designation. Only those contexts for which the user has authoring privileges will appear in the pull-down menu. As indicated in Figure 3.4-4, the Designations dialog box also allows you to remove designations from selected contexts.

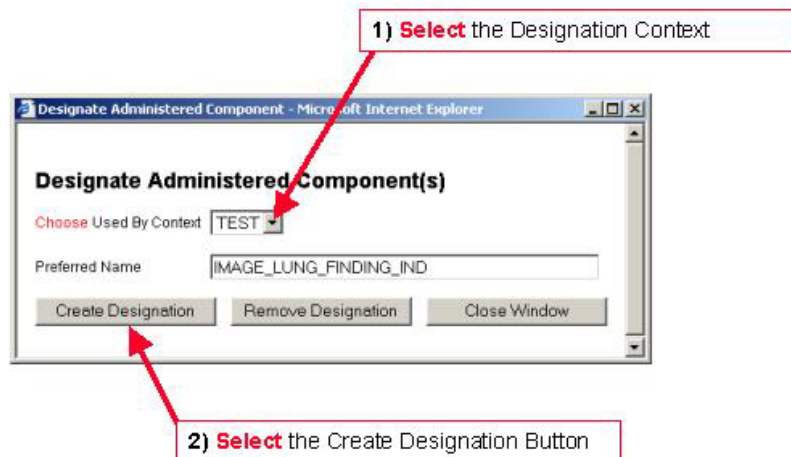


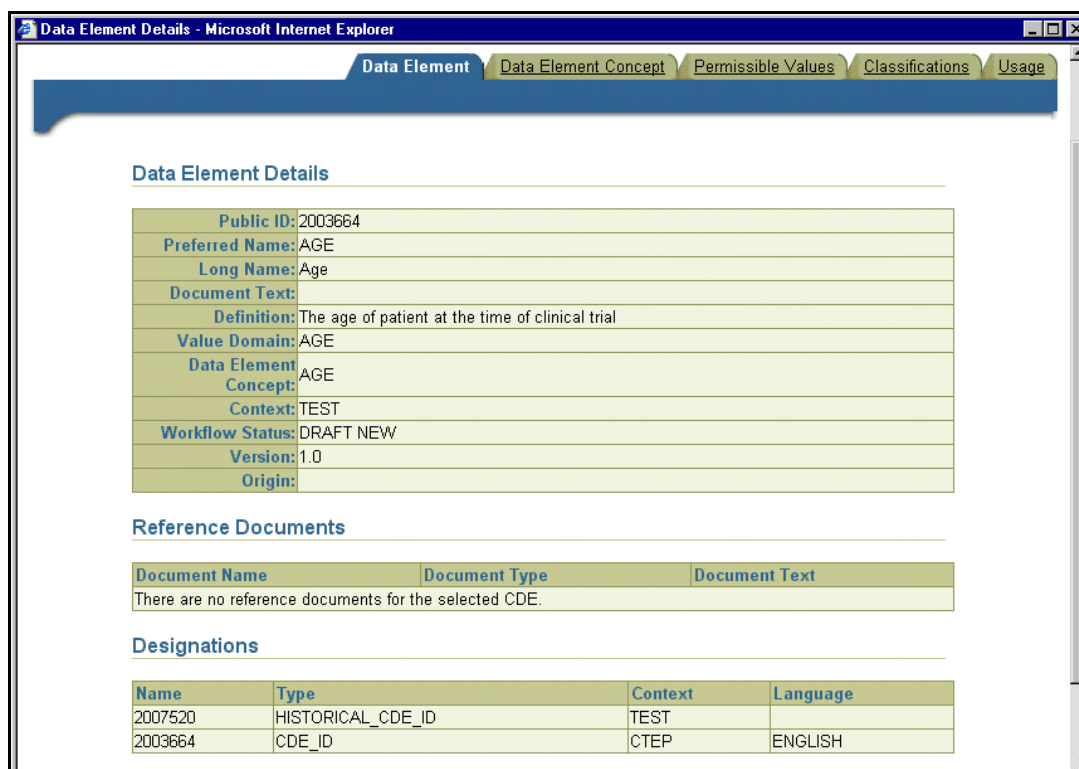
Figure 3.4-4 Pop-up Dialog for Creating and Removing Designations

If multiple rows are selected in the results table, the **Designations** button allows users to create “block designations.” In this case, the **Remove Designations** button will be disabled and the

Preferred Name field will be blank in the pop-up dialog box. A block designation will apply the indicated designation of a single context to all of the Administered Components in the selected block.

Designating a previously defined Administered Component is a very good alternative to creating a new component *ab initio*. Designating an Administered Component to additional Contexts provides the designator with READ ONLY privileges for that component. Changes to the Administered Component will be the responsibility of the “owning” context.

The third button in the Data Element results menubar is labeled **Details**. Clicking on the **Details** button brings up a Details page for that component, with the format of the page depending on the type of component. Different aspects of the component’s details are accessed via the folder tabs running across the top of the screen. Figure 3.4-5 shows the details page for a Data Element, with the tabs providing access to the component’s Data Element Concept, Permissible Values, Classifications, and Usage.



**Figure 3.4-5 The Details Page for a Data Element**

Details can only be viewed for one component at a time. Selecting more than one row in the results table disables the details option, and only one details window can be open at any given time. Keeping the Details window open for one component while subsequently selecting a new row in the table for detail viewing will repopulate the open Details page with information for that second component.

The fourth button in the Data Element results menubar, **Append**, is used to combine the results obtained from multiple consecutive searches. To use this option:

- Perform a search and select the records to keep.
- Press the **Append** button. This will remove all records that are not checked.
- Perform another search on the *same type* of Administered Component.
- Repeat these steps until you have all of the records you are interested in.

With each new search, only those rows that are selected will be appended to the growing result set. The append operation can only be applied to results of the same type. For example, you cannot append the search results from a Value Domain search to the results obtained by searching for Data Elements. The append operation will be applied to all subsequent searches until one of the following actions is taken:

- the **Clear Records** or top-level **Search** menu button is pressed; or
- a different option is selected, in either the **Search For** or **Search In** sidebar options.

Following the **Append** button is the **Show Selected Rows** option. Pressing this button simply removes all *unselected* rows from the table. The next button over is called **Get Associated**. Like the **Details** button, the **Get Associated** button is only enabled when a single row is selected.

This option allows you to access other types of administered components that are associated with the currently selected element. This operation cannot be applied to Data Element or CRF Question results, but is available for all other types of administered components. The associations that are available are:

- Access to the associated Data Element is provided from Data Element Concept, Value Domain, Classification Scheme Items, and Permissible Value search results.
- Access to the associated Data Element Concept is provided from Conceptual Domain search results.
- Access to the associated Value Domain is provided from Conceptual Domain and Permissible Value search results.

The last button is called **Clear Records** and is always enabled when results are displayed. This action will clear the results table and reset all of the sidebar options to their default settings.

The results menubar for Data Elements contains the largest number of options; all of the menubars associated with the other Administered Components are reduced versions of the Data Elements menubar.

## Creating and Editing New Administered Components

The CDE Curation Tool allows you to create and/or edit three types of Administered Components: Data Elements, Data Element Concepts, and Value Domains. The Edit screens are very similar to the Create screens. One difference is that in the latter case, the fields are either blank or contain default values. Edits screens also display the uneditable Public ID of the Administered Component.

The creation of a new Administered Component is initialized by pressing on the **Create** button in the dark blue navigation bar. This activates a pull-down menu allowing the user to select the type of component to create and the manner in which it will be created. Options include: (1) creating the component from scratch; (2) deriving a new component from a previously defined component; and (3) creating a new version of a previously defined component. The last two cases are very much like editing an existing component, as all of the mandatory attributes will have predefined values.



The following steps outline the process for creating a new component:

1. Select: **Create** → **<component type>** → **New** on the navigation bar.
2. Specify values for all mandatory attributes. Figure 3.4-6 shows the Data Element Creation form that is generated as a result of step 1. This form follows a standard layout for selecting and creating Administered Component attributes, with a clear demarcation between mandatory versus optional attributes.
3. Specify values for any desired optional attributes, using the same component creation form.
4. Validate the new component using the **Validate** buttons on the form.
5. Submit the component to the repository using the **Submit** button on the form.

This sequence of steps must be followed in creating any type of Administered Component. The only differences are in the mandatory and optional attributes that must be defined for the component. While these differences are reflected in the Creation form's content, the basic layout, widgets, and mechanisms are uniformly defined on all forms.

The following attributes are mandatory and must be specified for all component types:

1. **Context:** The first attribute on all forms is the Context for the component. A pull-down list allows you to select a context from a list of those for which you have authoring permissions.
2. **Name:** All components must be named. The ISO/IEC 11179 convention is to derive a new component's name from its defined subcomponents. This step, then, requires a multi-step specification. First you must select subcomponents that will serve as the building blocks for the new component. Second, you must derive the new component's name from the names of those subcomponents.

For example, a Data Element is composed of a Data Element Concept and a Value Domain, so the form in Figure 3.4-6 has mandatory slots for selecting these subcomponents prior to specifying the component's name. The next two slots allow users to modify the long and preferred names that will be auto-generated by the Curation Tool. The Long Name will contain the full names of the selected building blocks. The Preferred Name will contain the first four letters of the selected building blocks separated by underscores ("\_").

3. **Definition.** The Curation Tool provides three options for creating definitions. First, you can use the auto-generated default definition; each subcomponent definition is appended to the previous one's with an underscore. The definition is composed of the selected Data Element Concept's and Value Domain's definition. Second, you can search for and select definitions by clicking the **Search** button. Third, you can simply enter your own text.
4. **Version.** The (editable) default version for a new component is 1.0.
5. **Workflow Status.** The Workflow Status must be selected from the pull-down list associated with this attribute.<sup>10</sup>

---

<sup>10</sup> Workflow statuses are listed and defined on the caDSR Home Page under the [Context](#) section

6. **Effective Begin Date.** The date the element becomes public can be specified by directly typing the date in the textbox provided (formatted as MM/DD/YYYY) or by selecting a date via the calendar icon .

**NATIONAL CANCER INSTITUTE** **CADSR** **CDE Curation Tool**

Context: User Name : docu\_user

[Search](#) [Create](#) [Edit](#) [Options](#) [Help](#) [Logout](#)

[Validate](#) [Clear](#)

**Create New Data Element - \*Mandatory Attributes**

- 1) **Select** \*Context
- 2) **Select** \*Data Element Concept Long Name [Search](#) [Create New](#)
- 3) **Select** \*Value Domain Long Name [Search](#) [Create New](#)
- 4) **Verify** \*DE Long Name   Character Count (Maximum = 255)
- 5) **Verify** \*DE Preferred Name   Character Count (Maximum = 30)
- 6) **Create/Search** for \*Definition
- 7) **Enter** \*Version
- 8) **Select** \*Workflow Status
- 9) **Enter** \*Effective Begin Date   MM/DD/YYYY
- 10) [Validate](#) the New Data Element

**Create New Data Element - Optional Attributes**

- 11) **Create** Document Text
- 12) **Select** Classification Schemes/Classification Scheme Items

Selected Classification Schemes and Associated Items	
Classification Scheme	Items
<input type="text"/>	<input type="text"/>

- 13) **Select** Data Element Origin
- 14) **Select** Language
- 15) **Enter** Effective End Date   MM/DD/YYYY
- 16) [Validate](#) the New Data Element

**Figure 3.4-6 The Create Data Element Form**

As noted, all of the above attributes are mandatory for all components. The required building blocks vary with the component type. A Data Element Concepts is composed of an Object Class and a Property, along with Qualifiers for those components. A Value Domain is perhaps the most

complex to create, as it is composed of an Object Class, a Property, and a Representation, along with Qualifiers for each of these components.

In addition, both Data Element Concepts and Value Domains require that a Conceptual Domain is specified, and *Enumerated* Value Domains must specify Permissible Values. While it is not possible to create Permissible values from the main menubar, these can be created—as needed—during the process of creating a Value Domain. As to be expected, the *optional* attributes included on the component-specific forms also vary with the component types. Detailed step-by-step instructions for creating each of these component types is included in the online Help tool.

## Editing Administered Components

Editing these three types of administered components is very much like creating them, as the same forms are used for *modifying* attribute values. The only difference is that the attribute values are already filled in and must just be modified. And, as mentioned, the creation of a new component based on a previously defined component reduces to a special case of editing, as does the creation of a new version of a component.

To edit a component, begin by performing a search for that component type using the CDE Curation Tool’s search interface. Selecting a single component in the results table will enable the **Edit Selection** button, and clicking on that button will bring up the Edit screen.

It is also possible to perform “block edits” when two or more results are selected from the table. In this case the **Edit Block** button will be enabled, and the Edit screen for modifying a block of components will be generated. In editing groups of components, only a subset of the component attributes can be modified. Table 3.4-2 lists the attributes that can be edited for each component type when doing a block edit.

**Table 3.4-2 Component-Specific Attributes Available for Block Editing**

<u>Data Element</u>	<u>Data Element Concept</u>	<u>Value Domain</u>
Version (point or whole)	Version (point or whole)	Version (point or whole)
Workflow Status	Workflow Status	Workflow Status
Language	Language	Language
Origin	Origin	Origin
Effective Begin/End Dates	Effective Begin/End Dates	Effective Begin/End Dates
DEC Long Name	Object Class	Rep Qualifier
VD Long Name	Object Qualifier	Rep Term
Document Text	Property	Conceptual Domain
Classification Schemes	Property Qualifier	Data Type
Classification Scheme Items	Conceptual Domain	Unit of Measure
		Unit of Measure Display
		Format
		Minimum/Maximum Length
		High/Low Values
		Decimal Place

### 3.5 The caDSR Admin Tool

The caDSR Admin Tool is the main administrative interface to all of the caDSR features and components, and is intended for use by both Context and central caDSR administrators. A number of administrative and curatorial tasks are not supported in the CDE Curation Tool, and can be performed only through the caDSR Admin Tool. Access to this tool requires an administrative account.

Like the CDE Curation Tool, the caDSR Admin Tool provides a web interface for browsing, maintaining, and editing the administered components. The home page for this web site (Figure 3.5-1) consists of four panels:

- **Metadata Browsing and Maintenance** provides an interface for maintaining data elements, data element concepts, objects, properties, value domains, representations, conceptual domains, and classification schemes;
- **Submissions/Registrations and System Administration** is provided for tasks such as the creation of new user accounts, user groups, contexts, and workflow transitions;
- **Compliance Review Process** allows administrators to review, approve, and/or remove submitted case report forms.
- **Protocol/Form/Template Browsing and Maintenance** supports the browsing and maintenance of forms and templates in the registry.



Figure 3.5-1 The caDSR Admin Tool Welcome Screen

While the caDSR has functionality for the registration process, this capability is not currently used, and is included in this discussion for completeness only. The remainder of this section focuses on the Metadata Browsing and Maintenance capabilities.

### 3.5.1 The caDSR Search Interfaces

Clicking on any of the *Browsing and Maintenance* options in the top panel brings up a screen providing search and editing tools for that administered component type. The screen shot in Figure 3.5-2 shows the display that appears after clicking the Browse/Maintain option for Representation. As with all of the administered components, the search screen for a representation element includes:

- A pull-down menu for specifying the **Search Field**;
- A textbox for entering keywords to **Search For**;
- A **Context** specification area;
- A **Workflow Status** area;
- A textbox for entering keywords in the component's **Definition**; and
- A radio button for selecting the **Version** (latest only or all versions).

The **Search Fields** option specifies which search fields in the component should be matched against the keywords entered in the **Search For** textbox. Multiple search fields can be selected by holding down the **Ctrl** key and clicking the left mouse button.

A general rule is that all textboxes allow wildcard matching using the '%' sign. Thus, selecting "All Names" as the search fields, and entering "A%" in the **Search For** textbox, will find all components whose names begin with the letter "A." Similarly, entering the string "%vision%" in the **Definition** textbox will find all components that contain the word "vision" in their definition.

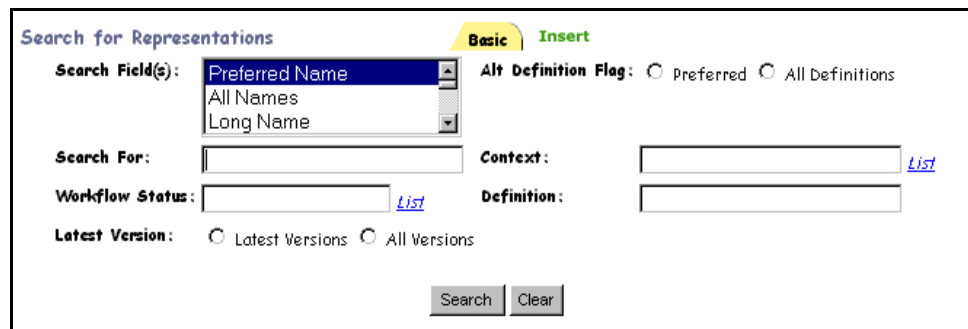


Figure 3.5-2 The Browse/Maintain Screen Associated With Representations

Both the **Context** and **Workflow Status** boxes require that you select the expression to be entered there from a picklist that appears when you click on the [List](#) icon. Finally, the default of searching only the latest version can be reset to search all versions of the data registry, using the radio button at the bottom of the screen.

Other options on the search screen will vary with the type of component. In addition to the options described above, a radio button at the top of the Representations search screen allows the user to set a flag specifying that either preferred definitions or all definitions should be used in

the search. Table 3.5-1 defines the additional search criteria options that may appear on different search screens.

**Table 3.5-1 Additional Search Criteria for Different Components**

<u>Search Criterion</u>	<u>Widget type</u>	<u>Function</u>
Context Use	Pull-down	Specifies that the component must be <i>Owned by</i> and/or <i>Used by</i> the selected Context
Data Element Concept	Picklist	Filters Data Elements by Data Element Concept
Value Domain	Picklist	Filters Data Elements by Value Domain
Conceptual Domain	Picklist	Filters Data Element Concepts by Conceptual Domains
Definition Flag <sup>11</sup>	Radio button	Specifies that all definitions or only preferred definitions should be used in the search
Type of Definition	Textbox	Specifies the type of definition sought. <sup>12</sup>
Domain Type	Radio button	Specifies that a Value Domain is enumerated or non-enumerated
Unit of Measure	Picklist	Filters Value Domains by unit of measure
Character Set	Picklist	Filters Value Domains by character set
Data Type	Picklist	Filters Value Domains by data type
Format	Picklist	Filters Value Domains by character set
Value Meaning	Picklist	Filters Conceptual Domains by Value Meaning
Class Scheme Type	Picklist	Filters Classification Schemes by Classification Scheme Type

In summary, four types of selection widgets are used: textboxes, radio buttons, pull-down menus, and picklists. All of the textboxes allow wildcard matching using the percent (“%”) character. Most of the pull-down menus allow multiple selections using the **Ctrl** key.

Picklists do not appear on the primary display; they are instead generated in pop-up windows when you click on the associated [List](#) icons. For short lists, the picklist is auto-generated with no further input from the user. Longer lists involving a very large number of possible values provide a textbox for the user to enter a text string that can be used to reduce the number of choices. Selecting a value from the picklist will automatically close the pop-up window and cause that value to appear in the associated option’s box on the primary screen. Clicking on the **Close** button without selecting a value will leave the option unspecified.

<sup>11</sup> Alternatively called “Alt. Definition Flag” and “Definition Search” on different screens.

<sup>12</sup> The ISO model provides for the ability to handle multiple definitions for one administered component, thus creating the need for multiple definitions types. Currently, the caDSR supports only the “Primary Description” definition type.

As noted above, with the exception of the Definition Flag at the top of the screen in Figure 3.5-1, all of the search criteria options shown there are included on all of the search screens. Table 3.5-2 maps the additional options described in Table 3.5-1 to the component-specific search screens on which they appear.

**Table 3.5-2 Additional Search Criteria Mapped to Specific Components**

	Data Element	Data Element Concept	Object Class	Property	Value Domain	Representation	Conceptual Domain	Classification Scheme
Context Use	✓	✓			✓			
Data Element Concept	✓							
Value Domain	✓							
Conceptual Domain		✓						
Definition Flag			✓	✓		✓	✓	✓
Type of Definition				✓				
Domain Type					✓			
Unit of Measure					✓			
Character Set					✓			
Data Type					✓			
Format					✓			
Value Meaning							✓	
Class Scheme Type								✓

## Basic Search

All of the Browsing and Maintenance screens support basic search, with interfaces similar to that shown in Figure 3.5-1. Basic search allows the user to search for all components by name, definition, context, workflow status, and/or version.

The first name given to an Administered Component is stored as its preferred name. When searching by name, candidate matches are ranked differently depending on whether the user has selected the “Preferred Name,” “All Names,” or “Long Name” options. If “Preferred Name” is selected, then the keywords entered in the **Search For** field will only be matched against preferred names.

Several other fields defined in the **Search Fields** pull-down menu specify various Ids by which the component may be accessed. These include Public ID, Historical\_CDE\_ID, NCI\_Concept\_Code, UMLS\_CUI, TEMP\_CUI, etc. Using the “All Names” option will include all of these fields. And, as with most of the Admin Tool’s pull-down menus, multiple fields can be selected by holding down the **Ctrl** key while simultaneously selecting the desired fields.

As with preferred names, the first definition given to an Administered Component is also recorded as the *preferred* definition; selecting the **Preferred Definition** radio button will limit the

matching to preferred definitions only. The remaining search fields common to all of the basic search screens are:

- **Workflow Status:** This is the administrative (workflow) status of an Administered Component (such as Draft, Reviewed, Approved, Released). Select the *List* icon to get the list of values.
- **Context:** This is the context of the Administered Component's preferred name (i.e. the first Context associated with the component).
- **Latest Version or All Versions:** The caDSR allows you to retrieve just the latest version of an administered component or all versions of that component. Each instance of an Administered Component is uniquely identified by its preferred Name, preferred Context, and Version.

## Full Text Search

In addition to basic search, the search screens for Data Elements, Value Domains, and Concept Domains also support *full-text* search. Full-text searching allows you to enter unstructured, full-text information about a given administered component. This full text combines the name, long name, description fields, and other relevant fields. If there is a hit on the search term in any of the combined text, the administered component is returned. To switch to full-text search mode, click on the *Full Text* tab included on the menu bar running across the top of the search screen for one of these components. Figure 3.5-2 shows the full-text search screen.

Full-text searches are case insensitive and support wildcard (%) matching. Specific search options are provided via the pull-down menus surrounding the textboxes, and up to three text strings can be specified. These strings can be combined differently according to the selected Boolean operator.

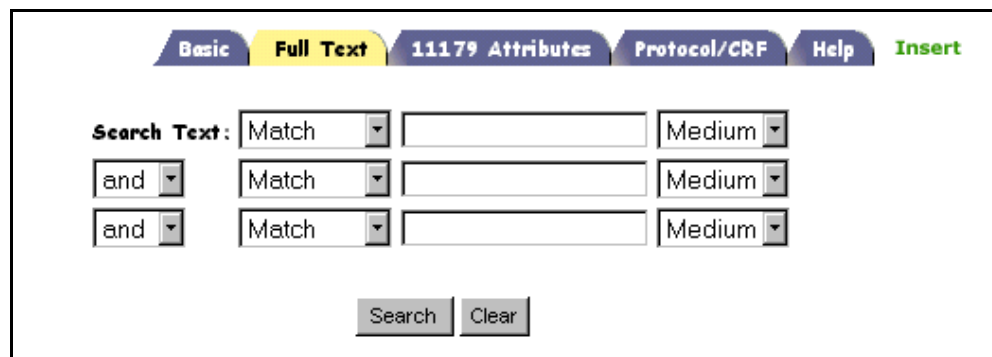


Figure 3.5-2 Full-Text Search Screen

The leftmost pull-down menus specify the Boolean operators to be used in combining the expressions occurring in the preceding and current textboxes. The choices include *and*, *or*, *not*, and *near*, where:

- *and* evaluates to TRUE if the candidate component matches both expressions;
- *or* evaluates to TRUE if the candidate component matches either expression;
- *not* evaluates to TRUE if the candidate component matches the first expression but does *not* match the second expression;



- *near* evaluates to TRUE if the first expression occurs “near” the second expression in the candidate component’s matching terms.

The Boolean operators are evaluated in top-to-bottom order, and there is no precedence to the operators other than that defined by position.

The next pull-down menu (“Match”) allows the user to constrain *how* the matching is done; options include *Match*, *Starts with*, *Stem*, and *Soundex*. The default, *Match*, specifies that unconstrained lexicographic matching should be used. The other options are as follows:

- *Starts with* searches for text that starts with the expression provided in the textbox;
- *Stem* searches for text that shares the same stem as the textbox expression;
- *Soundex* searches for text that “approximately sounds” like the search expression (e.g. “reed” and “read,” “wage” and “age,” etc.). Soundex matching is based on a very simple encoding which approximately captures the significant phonemes in a string.<sup>13</sup>

Finally, each search condition can also be prioritized as *High*, *Medium* (the default), or *Low*. Hits on high priority expressions will be returned first, followed by medium, and then low priority expressions. Only those components whose text satisfies all of the search criteria (as combined by the selected Boolean operators) are returned by the full-text search.

### 11179 Attribute Search

The search interface for Data Elements provides one additional mode of search—the *11179 Attribute Search*. The 11179 Attribute search screen (Figure 3.5-3) allows the user to search by the ISO/IEC 11179 attributes in four categories: *Naming and Identification*, *Definitional*, *Representational*, and *Administrative*.

Figure 3.5-3 The 11179 Attribute Search Screen

The first two categories simply provide subsets of the search fields available with the basic search interface. Figure 3.5-3 shows the interface when the third tab, *Representational*, has been selected. The Representational interface introduces additional search criteria for data elements, allowing the search to drill down into the data element’s value domain attributes without explicitly using the Value Domain’s search interface.

<sup>13</sup> The Soundex matching algorithm was first used by the United States Census Bureau. Additional information can be found at [http://www.archives.gov/research\\_room/genealogy/census/soundex.html](http://www.archives.gov/research_room/genealogy/census/soundex.html).

The last tab, *Administrative*, allows the user to search for Data Elements by their administrative attributes, including the element’s registration status, the submitter, the steward, and the registrar authority, as well as the organizations associated with these individuals.

### 3.5.2 Working with the Search Results







Search results are returned according to the specified search criteria and the individual user’s privileges, with the results listed in a table below the search form. Just as the search criteria options vary with the type of Administered Component being searched for, the attributes shown in the results table vary according to the properties that are relevant to that component type.

Several fields however, are included in all search result tables. The leftmost column always displays a *Browse* icon in the form of a magnifying glass. Clicking on that icon brings up the Details page for the administered component in that row of the table.

If the user has editing permissions for that record, then the second column displays a *Modify* icon in the form of a pencil. Alternatively, if the user does not have editing privileges, the second column is empty. Note that if the user has neither browse nor edit permissions on a given component, that record will not be displayed in the results table—even if its attributes otherwise matched the search criteria.

The remaining columns in the results table correspond to the fields in the component type’s search form. Thus, for a data element, the fields are: *Name*, *Name Type*, *Preferred Name*, *Owned by*, *Used By*, *Version*, *Long Name*, *Workflow Status*, *Data Element Concept*, *Definition*, *Type of Definition*, and *Value Domain*. Figure 3.5-4 shows a part of the results table obtained by a search for Data Elements in the “Test” Context whose names begin with the letter “a.”

Large search result tables have navigation buttons at the bottom of the list. If the result set is large, you may need to scroll through sets of records using the **Next**, **Previous**, **First**, and **Last** buttons. The **ReQuery** button re-executes the same query; the **Count** button shows the total number of records that met the search criteria. A user may execute a new search by pressing **Clear**, entering new search criteria, and pressing **Search** again.

Browse	Modify	Name	Name Type	Preferred Name	Owned by	Used by	Version	Long Name
		ADDRESS_Name	Preferred Name	ADDRESS_Name	TEST		1	Address Name
		AGE	Preferred Name	AGE	TEST		1	Age
		AGENT	Preferred Name	AGENT	TEST		4	Agent

**Figure 3.5-4 A Part of the Results Table for a Data Element Search**

Figure 3.5-5 shows the Details screen for the AGE data element. The *Permissible Values* tab has been selected, and the main frame shows the table of permissible values for an element of this type. The screen also provides access to information about the *Value Domain*, *Data Element Concept*, and any *Documents* associated with this element. Each node in the tree-structured display on the left is selectable and provides an alternate way of navigating through the information associated with the data element.

The screenshot displays a web application interface for managing data elements. At the top, there are tabs: 'Data Element', 'Value Domain', 'Permissible Values', 'Data Element Concept', 'Documents', and 'Help'. The 'Value Domain' tab is active.

On the left, a tree view shows the hierarchy: 'Data Element' > 'AGE' > 'Value Domain'. Other options in the tree include 'Examples', 'Alternate Names', 'Alternate Definitions', 'Admin Notes', 'Description', 'Detail Description', 'Comments', 'Documents', 'Related Data Elements', 'Admin Infos', 'Contact', 'Derived Data Elements', and 'Classification Schemes'.

The main content area is titled 'Search for Values'. It contains two input fields: 'Value:' and 'Value Meaning:', each followed by a 'List' link. Below these fields are 'Search' and 'Clear' buttons.

Below the search section is a 'Values Information' section, which includes a table with the following data:

Value	Value Meaning	Description	Effective Begin Date	Effective End Date	Low Value	High Value
65-69 yrs.	65-69 YRS.		07/08/2002			
70-80 yrs	70-80 YRS		07/08/2002			
> 80 yrs	> 80 YRS		07/08/2002			

Below the table, it says 'Records 1 to 3 of 3' and there is a 'ReQuery' button.

Figure 3.5-5 Details Page for the AGE Data Element

### Summary of Search Screen Behaviors

- The search criteria fields are not case-sensitive.
- If search criteria are provided in multiple fields, a logical AND is applied – i.e., the matched component must satisfy all of the criteria.
- When no criteria are specified, the search retrieves all records for that component type to which the user has access.
- The percent sign (%) may be used for wildcard matching.
- [List](#) icons allow the user to choose from a list of valid values for the given field.
- The **Search** button invokes the actual search;
- The **Clear** button resets all fields in the search form;

### 3.5.3 Maintenance Screens for Administered Components

Maintenance screens combine the appearance of the browser screens with the functionality of the displays used to create new components. Figure 3.5-6 shows the Maintenance screen for a Data Element named COORDINATING\_GRP\_PROTOCOL\_NUM.

The Maintenance screen is reached by clicking the pencil icon displayed in the second column of the results table. Editing capabilities include adding, updating, and deleting the information fields shown in the Maintenance screen. The tree icons in the left panel provide access to editing screens for the corresponding attributes of the component. The Maintenance screen for a particular Administered Component is only accessible to users having *Update* and *Delete* privileges.

**Data Element**   **Value Domain**   **Valid Values**   **Data Concept**

**Data Element**   **Maintain Data Element**

**\* Name:** COORDINATING\_GRP\_PRO

**Long Name:** Coordinating Group Protoc

**\* Definition:** The numeric or alphanumeric identification assigned to the

**\* Context:** CTEP [List](#)

**\* Value Domain(VB):** PRIM\_SITE [List](#) [New](#)

**\* Data Concept(DEC):** COORDINATING\_GRP\_PRO [List](#) [New](#)

**Effective Begin Date:** [Dates](#)

**Effective End Date:** [Dates](#)

**Change Note:**

**\* WkFlow Status:** RELEASED [List](#)

**Version:** 2.3

**Latest Version:** Yes

[Apply](#) [Delete](#) [Undo](#) [Add New](#)

[Copy](#)

Figure 3.5-6 Maintenance Screen for Modifying a Data Element

### 3.5.4 Creating Administered Components

All of the component-specific search screens include a rightmost tab with the keyword *insert* displayed on it. Clicking on this tab for the Data Element search interface brings up the screen shown in Figure 3.5-7, which allows the user to define a new Data Element. Alternatively, from the Maintenance screen for an administered component (e.g., Figure 3.5-6), you can press the **Add New** button (bottom, right) to create a new component.

Figure 3.5-7 shows the interface for creating a new data element component. As some of a new component's attribute values may themselves be administered components (e.g., the Value Domain or Data Concept for a Data Element), the interface provides mechanisms for selecting from the set of currently defined components, as well as recursively creating a new component on the fly. The [New](#) icon allows you to (1) create the new component, and (2) subsequently assign it via the [List](#) icon.

**Data Element** [New Search](#)

**Data Element**

**Enter Data Element**

\* **Name:**

**Long Name:**

\* **Definition:**

\* **Context:**  [List](#)

\* **Value Domain(VD):**  [List](#) [New](#)

\* **Data Concept(DEC):**  [List](#) [New](#)

**Effective Begin Date:**  [Dates](#)

**Effective End Date:**  [Dates](#)

\* **WkFlow Status:**  [List](#)

**Note:** \* - Mandatory field

**Figure 3.5-7 Creating a New Data Element Component**

The caDSR requires that most of the mandatory attributes of ISO/IEC 11179 must be supplied. Exceptions are the submission, registration, and stewardship assignments, which are not included in the form. Required fields are indicated by an asterisk preceding the attribute name. If the value for a mandatory field is not currently known, the system will accept the designation “UNASSIGNED.”<sup>14</sup> Once the administered component has been created, the application will display the Maintenance screen, where you can add additional relationships and information for that component.

Three of the interfaces for creating new Administered Components also provide access to EVS terms and definitions. Figure 3.5-8 shows the interface for defining a new Property. Just to the right of the textbox for entering the Preferred Name are two hyperlinks providing access to (1) lists of EVS concepts, and (2) a submission form for proposing new terms. The caDSR Admin Tool does not link to an EVS navigation window, but instead provides a simple interface that is consistent with other windows in the tool’s interface. Clicking on the [EVS Concept List](#) link causes a small text box to appear (as in Figure 3.5-8) where you can enter search terms. After you enter the search terms and press **OK**, the list of concepts matched to those terms in the EVS will be generated in a new window. Selecting a concept name there closes that window and

<sup>14</sup> The ISO/IEC 11179 Compliancy Test will look for and warn users of mandatory fields with an entry of ‘UNASSIGNED’.

returns you to the Property Creation form with the selected concept now appearing in the **Preferred Name** slot. The **Long Name**, **Definition**, **Definition Source**, and **Database** fields will also be populated from the information provided by EVS. Similar capabilities are provided on the forms for the creation of object classes and representations.

The second hyperlink is provided for cases where the EVS search fails to find matching concepts. Clicking on **Notify EVS** generates a submission form that can be used to notify EVS that the concept sought for could not be found.

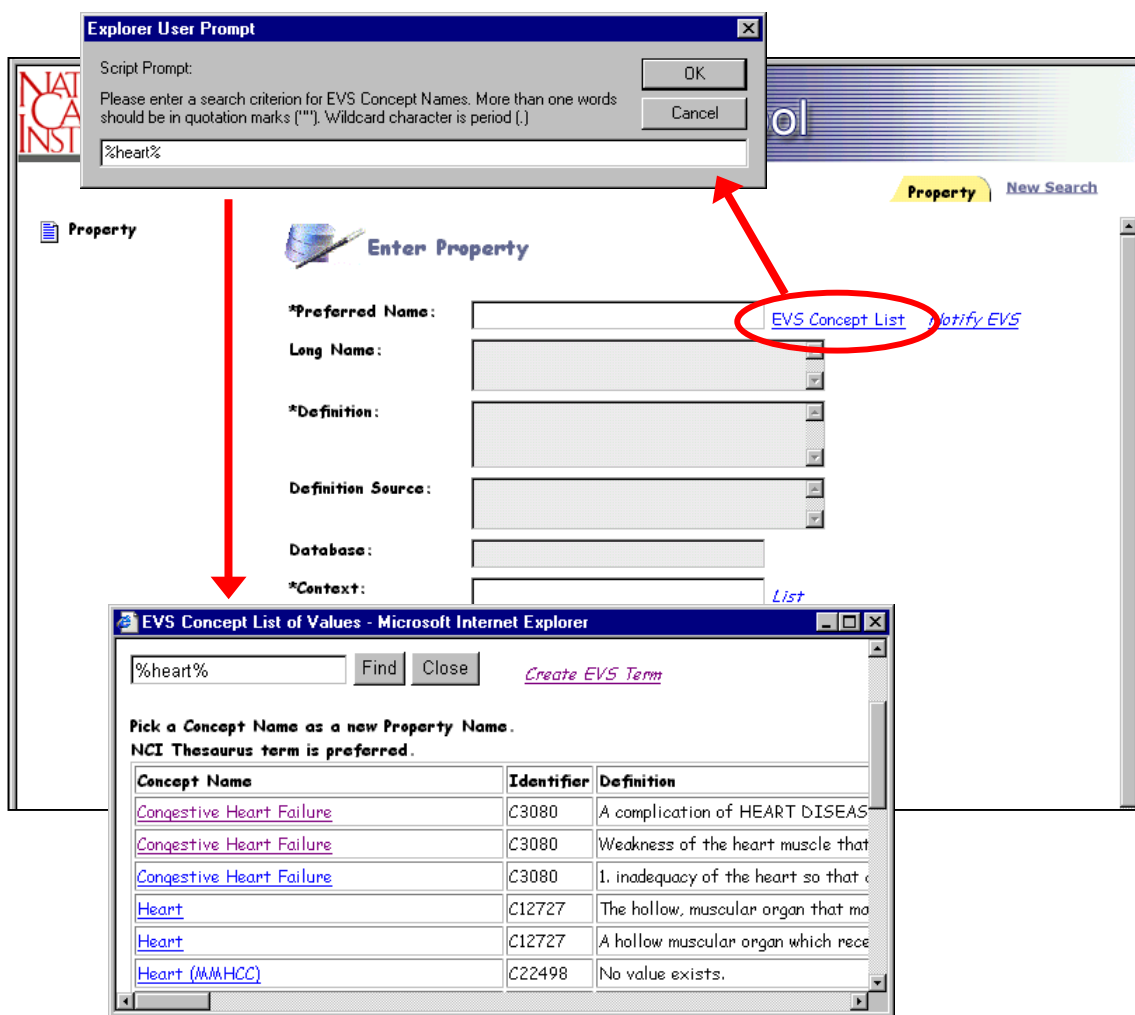


Figure 3.5-8 Accessing EVS Terms and Definitions

### 3.6 Comparison of the caDSR Tools

The caDSR Tools suite provides different capabilities to different users, depending on the permissions that have been granted as well as on the defined needs of those user communities. As a result, there is some unavoidable redundancy in the capabilities provided by the various tools. Each tool, however, provides some set of non-redundant functionality to the groups using that tool. Table 3.6-1 summarizes these capabilities and is also available on the web at <http://ncicb.nci.nih.gov/NCICB/core/caDSR/Matrix.xls>.

**Table 3.6-1 Capabilities Served by the caDSR Tools**

	CDE Browser	Compliance Review	CDE Curation	Compliance Review Response	caDSR Admin	UML Loader	CRF Loader
<b>Search/View</b>							
Search for All Administered Components <sup>15</sup>	✓				✓		
View All Details for Administered Components					✓		
View All Details for Administered Components related to Data Elements	✓				✓		
Search for DE, DEC, Val D	✓		✓		✓		
Search Protocols and CRFs				✓	✓		
View Protocol CRFs that are ready for review				✓			
Assign CRF to a Protocol					✓		
<b>Determine CRF Compliance</b>							
Select CRF for Review				✓	✓		
Match CDEs to CRF Questions		✓					
Automatically Match CDEs to CRF Questions		✓					
Identify Questions that should become New Draft CDEs		✓					
Load DE Information into caDSR Extension Tables <sup>16</sup>		✓					✓
Update DE Information in caDSR Extension Tables				✓			
Search for DE and Valid Values		✓					
Search for CRF Questions [from a CRT review]			✓				
Recommend CDEs for Use in CRFs		✓					
Communicate CRF Compliance Status and Comments (* via comments)		✓*		✓			
Respond to CRF Review (* via comments)		✓*		✓			
Delete CRF					✓		
<b>Create/Edit</b>							
Create and Edit All Administered Components					✓		
Help User Create ISO/IEC Compliant DE, DEC, Val D			✓				
Create DE, DEC, Val D Based on Existing DE, DEC, Val D			✓				
Create DE, Permissible Values				✓			
Create Permissible Values and Value Meanings (* Using EVS search)			✓*		✓		
Create New Versions of All Administered Components (* DE, DEC, VD only)			✓*		✓		
Create New Versions of CD, DE, DEC, Val D					✓		
Manually Apply "point" vs "whole" Versioning Rules			✓				
Create New Versions of DE, DEC, Val D			✓		✓		

<sup>15</sup> For all entries in Table 3.6-1, the term “Administered Component” includes Data Element, Data Element Concept, Value Domain, Conceptual Domain, Object Class, Property, and Representation. elements.

<sup>16</sup> Extension Tables include Protocol, Case Report Form , Question, and Value Domain

	CDE Browser	Compliance Review	CDE Curation	Compliance Review Response	caDSR Admin	UML Loader	CRF Loader
Find Definitions in EVS or suggest new terms and definitions			✓		✓		
Find Object Classes and Properties in EVS or suggest new terms			✓		✓		
Create/edit Protocols					✓		
Release Protocol for CCRR Review		✓					
Review Protocol CRFs		✓		✓			
Automatically generate preferred names and long names based on ISO 11179 naming guidelines			✓				
Automatically generate a DE Definition based upon the selected DEC and VD component definitions			✓				
Automatically generate a DEC Definition based on Object Class, Property and Representation terms			✓				
Automatically create new Object Class, Property and Representation terms as a byproduct of Creating DEC			✓				
Update workflow status (* w/Restrictions)		✓*	✓	✓*	✓		
Designate "Used By" Context							
Designate DE, DEC, Val D			✓		✓		
Apply Workflow Status Restrictions for Designation Rules			✓		✓		
Designate All Administered Components (* DE, DEC and VD only)			✓*		✓		
Import/Export							
Export CDEs to XML or MS Excel format	✓						
Import CRF data extracted into MS-Excel File Format							✓
Import Data Element Concepts and Data Elements from UML Class Diagram						✓	
View electronic CRF image file (e.g, MS Word, PDF, etc.)	✓				✓		
Security/User Access							
Requires User Account to access		✓	✓	✓	✓	✓	✓
Enforces Role based User access		✓	✓	✓	✓	✓	✓
System Administration							
Create/Edit User Account					✓		
Create/Edit User Group					✓		
Create/Edit Context					✓		
Create/Edit Lookup Lists of Values					✓		
Delete Administered Components from the caDSR database					✓		
Metadata Submission/Registration							
Assign Stewards to administered components					✓		
Submit administered components to Registration Authorities					✓		
Maintain administered component submissions and registries					✓		



## **Genome Analysis Tools**

## 4.0 BIOgopher

BIOgopher is a powerful ad hoc query and reporting tool that enables researchers to annotate entries in Microsoft Excel™ spreadsheets with data generated by the NCICB's Cancer Bioinformatics Infrastructure Objects. BIOgopher presents a web-based, graphical user interface with which a user can build complex queries incorporating data from any number of user-supplied local spreadsheets. The results of such queries are then delivered to the user as a new spreadsheet in which the originally submitted data and the caBIO data are merged.

As an example, suppose a scientist has a spreadsheet consisting of genomic sequence data, and that one of the columns in that spreadsheet contains the GenBank accession number for each row. A natural way of enhancing these data might be to incorporate the cellular pathway information associated with each of these accession numbers.

In most environments, this would require visiting another web site such as BioCarta, entering appropriate queries consecutively, and cutting and pasting each response back into the spreadsheet. BIOgopher allows the user to accomplish this task with a single query by indicating which column contains the accession numbers and specifying what pathway information should be included. The resulting spreadsheet returned by BIOgopher will contain a new column specifying the selected pathway information for each gene, in conjunction with the original columns.

BIOgopher also allows users to build queries interactively without requiring any initial spreadsheets. The search query can be assembled using pull-down menus to select caBIO domain objects such as Clone, Disease, Gene etc., from the caBIO “object tree.” After entering these query objects, the user then selects the desired data fields and values to be used in the query, and a spreadsheet containing the data extracted from the retrieved caBIO objects will be created. The caBIO domain objects provide an exhaustive model of the many biomedical and genomic entities commonly found in biomedical research today.

### 4.1 Getting Started

To begin using BIOgopher, open an Internet Explorer<sup>17</sup> window and visit the BIOgopher home page at <http://biogopher.nci.nih.gov/>. The first page you will see is the BIOgopher Welcome Page containing three folder tabs in the left-hand panel (see Figure 4.1-1):

1. The leftmost *local sources* tab is used for uploading spreadsheet files.
2. The center *search criteria* tab allows users to interactively enter search criteria.
3. The rightmost *report format* tab is used for specifying how the results will be formatted.

The right panel provides context-sensitive help. For example, when the first tab is selected, the right panel shows information about the local data sources that the user has defined and provides tips on defining these. Alternatively, if the *search criteria* tab (center) is selected, the right panel provides help on building and working with queries. Finally, when the rightmost tab is selected, instructions on formatting your results become available.

---

<sup>17</sup> The current version of BIOgopher supports Internet Explorer only.

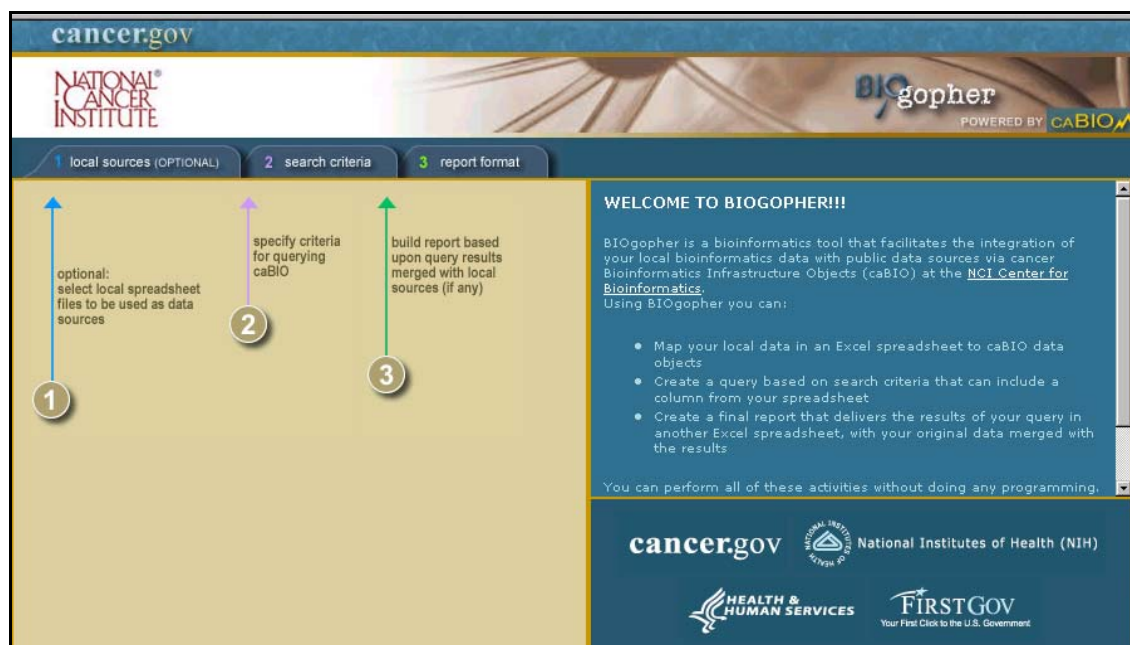


Figure 4.1-1 The BIOgopher Welcome Page.<sup>18</sup>

## 4.2 Example 1: Formulating Ad Hoc Queries

This first example demonstrates the formulation of ad hoc queries using the BIOgopher interface and the subsequent transfer of the retrieved data to a desktop spreadsheet. The two main tasks outlined in this example are: (1) formulating a query, and (2) formatting the report output.

To access the form for building a query, begin by selecting the center tab in the left-hand panel labeled *search criteria*. The screen should now look like the display in Figure 4.2-1. As indicated there, you can use this page to build a new query from scratch or to edit an existing query. The left panel holds a form for initializing a new query (or editing a previous one) and the right panel contains context-sensitive help on this topic.

The first question to answer in building a new query is: Which caBIO domain object best fits the query data? Clicking on the down arrow in the pull-down menu for slot A in Figure 4.2-1 displays a list of possible objects. To see the definitions and applications associated with these objects, you can click on the hyperlink in the prompt “Select the [object](#) for which you will build a query.” This will open a page listing the caBIO domain objects’ definitions, associations, and applications. In this example we will be searching for genes, so we select the Gene object from the pull-down menu.

BIOgopher allows you to maintain and work with multiple queries, and accordingly, requires that each query has a name that can be used to identify it. This example uses the name “GeneQuery.” Clicking the [Create](#) button initializes the query and takes you to the next screen (Figure 4.2-2), where more specific search criteria will be selected.

<sup>18</sup> The BIOgopher banner shown in Figure 4.1-1 is included on all of the BIOgopher screens; the remaining screen shots in this section crop the banner however, in order to save space and emphasize the pages’ contents.

Figure 4.2-1 Search Criteria Form

The right panel in Figure 4.2-2 is actually divided into two subpanels, with the upper half displaying context-sensitive help and the lower half displaying the partially formulated query. Thus far we have specified that our query will be retrieving Gene objects, as reflected by the solid white button labeled **Gene** in the query subpanel.

The left-hand panel in Figure 4.2-2 shows the attributes of a Gene object, along with related objects such as CMAPOntologies, Chromosome, and so forth. Mousing over any of these elements in the display generates descriptive pop-ups that document the attributes and objects. The pop-up in Figure 4.2-2 is displayed as a result of mousing over **Chromosome**.

Figure 4.2-2 Specifying Search Criteria: the Gene Object and its Attributes

These attributes and related objects can now be used to selectively filter out genes we are not interested in. In this example we will search for several genes by name, so we select the **Name/Symbol** attribute. Selecting an attribute in the left panel causes the context-sensitive help subpanel to be replaced with the form for specifying attribute values shown in Figure 4.2-3.

**Figure 4.2-3 Form for Specifying Attribute Values**

The six buttons to the right of the WORKING VALUES textbox work as follows:

- **add value** generates a pop-up textbox where the user can type in a new value
- **remove values** deletes all *selected* entries from WORKING VALUES; if no values are currently selected this action has no effect.
- **acquire values** imports values from a specified field in a local spreadsheet (see example 2).
- **browse caBIO** retrieves possible values for the selected attribute from the caBIO database.
- **update query** adds all of the entries in WORKING VALUES to the query as a Boolean disjunctive clause.
- **cancel** cancels the current operation and leaves the query unmodified.

Each button also has a question mark next to it; clicking on that icon will produce a pop-up describing that option. For this example we select the first button to manually enter the names of the following genes: TNF, EGFR. Note that each of these names must be entered separately—you can not use a comma-delimited entry to specify multiple values. After adding these values, clicking the **update query** button updates the query subpanel to:

**Figure 4.2-4 Updated Query Panel After Specifying Attribute Values**

As illustrated in Figure 4.2-4, the solid white button labeled **Gene** in the previous query subpanel now has a “–” in it, indicating that this is an “expanded” expression. Clicking on the button now will “collapse” the subterms specifying the attribute values, and change the sign on the button to a “+.”

A good deal of redundancy results from the curation of the same genes from multiple species. Thus, to reduce the size of the results table, we specify one additional attribute—the taxon—before generating a report. To do this, we scroll down to the bottom of the Gene object’s tree in the left panel, expand the **Taxon’s** node, and select **Scientific Name**. To now view the available scientific names, we select **browse caBIO**. Figure 4.2-5 shows the browser window that is generated in response.

**BROWSE TAXON - SCIENTIFIC NAME**

Done Select All

1 through 25 of 76

BEGIN | PREVIOUS | NEXT | END

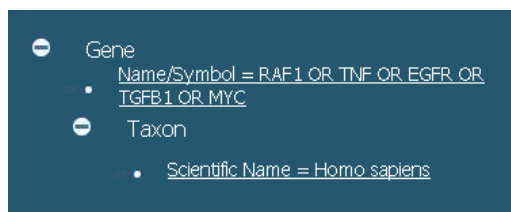
	ID	Scientific Name	Abbreviation	Common Name	Ethnicity Strain
<input type="checkbox"/>	6	Mus musculus	Mm		
<input type="checkbox"/>	7	Mus musculus domesticus	Mm		
<input type="checkbox"/>	8	Mus musculus	Mm		129
<input type="checkbox"/>	9	Mus musculus	Mm		129 - C57/B6 - FVB
<input type="checkbox"/>	10	Mus musculus	Mm		129/01a
<input type="checkbox"/>	11	Mus musculus	Mm		129/Sv
<input checked="" type="checkbox"/>	5	Homo sapiens	Hs		

	ID	Scientific Name	Abbreviation	Common Name	Ethnicity Strain	Is Preferred	
Deselect	5	Homo sapiens	Hs			true	<a href="#">At</a>

**Figure 4.2-5 Browse caBIO for scientific names of taxa**

Clicking the checkbox for “Homo sapiens” at the bottom of the upper panel results in that entry being added to the table in the lower panel. While we could continue to select additional entries, our goal is to limit the number of results, so we simply click **Done** to return to the previous page. There, “Homo sapiens” now appears in the WORKING VALUES box, so we press **update query** to add this constraint to our search criteria. Figure 4.2-6 shows the updated query.



**Figure 4.2-6 The updated query panel**



Having specified these search criteria, we can now generate a report by selecting the rightmost tab in the left-hand panel labeled “report format.” Note that the report form that is generated (Figure 4.2-7) allows you to specify a report format for *any* query—not just the one you are currently working with—as indicated by the pull-down menu in slot A. In this case however, we have just a single query to work with, “GeneQuery.”

Like queries, reports can be edited and reused, so our next task is to name the report we will be creating. For this example we enter the name “GeneReport” and click the **Create** button to initialize the report.

The screenshot shows the 'report format' tab in the BIOgopher interface. It is divided into two main sections: 'CREATE A NEW REPORT' and 'WORK WITH AN EXISTING REPORT'. The 'CREATE A NEW REPORT' section has three steps: A (Select the query for which you will build a report, with a dropdown menu showing 'GeneQuery'), B (Enter a name for this report, with a text input field containing 'GeneReport'), and C (Create the report, with a 'Create' button). Below this is a separator '- OR -'. The 'WORK WITH AN EXISTING REPORT' section has two steps: A (Select the report with which you would like to work, with a dropdown menu) and B (Work with the report, with 'Edit', 'Rename', and 'Remove' buttons). On the right side, there is a blue panel titled 'CREATING REPORTS' containing explanatory text about how reports are generated and how to use the interface.

**Figure 4.2-7 Initializing a report format**

The left-hand side of the screen that is generated (Figure 4.2-8) in response to pressing the **Create** button is similar to the query formulation screen. The left-hand panel again shows a tree-like structure under the Gene object, exposing its attributes and related objects. In this case however, selecting attributes and objects does not affect the query, but instead, affects what will be displayed in the results table.

To complete this example we select three immediate attributes of the Gene object—**Name/Symbol**, **OMIM ID**, and **Title**—along with the name attribute of the Chromosome object that the gene is associated with. The immediate attributes of the Gene are selected by simply clicking on the attribute name in the left panel. To add the chromosome name, we must first expand **Chromosome** by clicking on the button to the left of that object, and from the newly expanded subtree, selecting **Name**.

The right-hand panel in Figure 4.2-8 shows the currently defined output format. Each field in the report is labeled with an <object-name attribute-name> tag, as listed in the box in the upper right panel. The scrollable lower right panel also lists these fields in the tabulated format in which will appear in the report. Note that each panel in the BIOgopher display is resizable—Figure 4.2-8 was generated by clicking the left mouse button and dragging the Report Format panel to the left to expose all four columns.

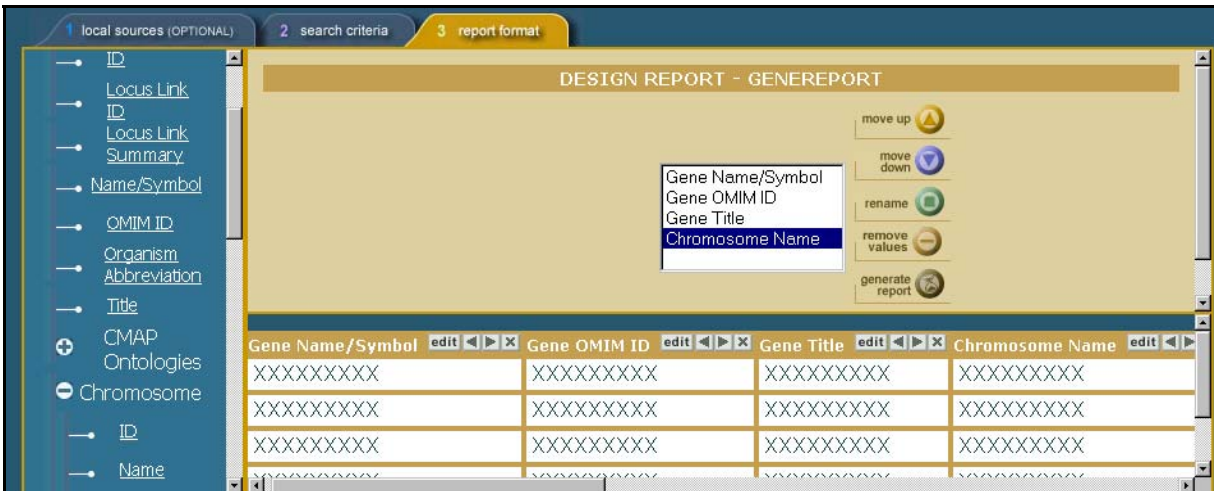


Figure 4.2-8 Defining the output fields

The output can be further modified by working in the Report Format panel. Each column contains left and right arrows which will move the selected column in the indicated direction. Columns can also be renamed using the Edit option, or deleted, using the **X** icon. The same editing actions are also available in the upper right-hand panel. To simplify the output, we rename these fields to “Gene,” “OMIM ID,” “Chromosome,” and “Title.” Also, because the “Title” field is likely to scroll off the page, we move this field to the last column.

Clicking **generate report** then produces the results in Figure 4.2-9. The formatted results are displayed as an HTML page that can be downloaded to a local Excel file using the hyperlink in **Click here to download**. Save this file as *geneSearch.xls* now, as we will use it as input in the next example.

CLICK [HERE](#) TO DOWNLOAD.

1 through 2 of 2

BEGIN | PREVIOUS | NEXT | END

Gene	OMIM ID	Chromosome	Title
EGFR	131550	7	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b)
TNF	191160	6	tumor necrosis factor (TNF superfamily, member 2)

Figure 4.2-9 The Results Screen

### 4.3 Example 2: Using Local Spreadsheets to Acquire Values

Assuming that you saved the output from the result set in Figure 4.2-9 as an excel spreadsheet, we can now use this as input to acquire values. To begin, select the leftmost tab in the left panel, **local sources**. This will bring up a new panel (Figure 4.3-1) with slots for specifying the file location, worksheet number, header row information, and row number where the data actually starts.



**Figure 4.3-1 The Spreadsheet Upload Page**

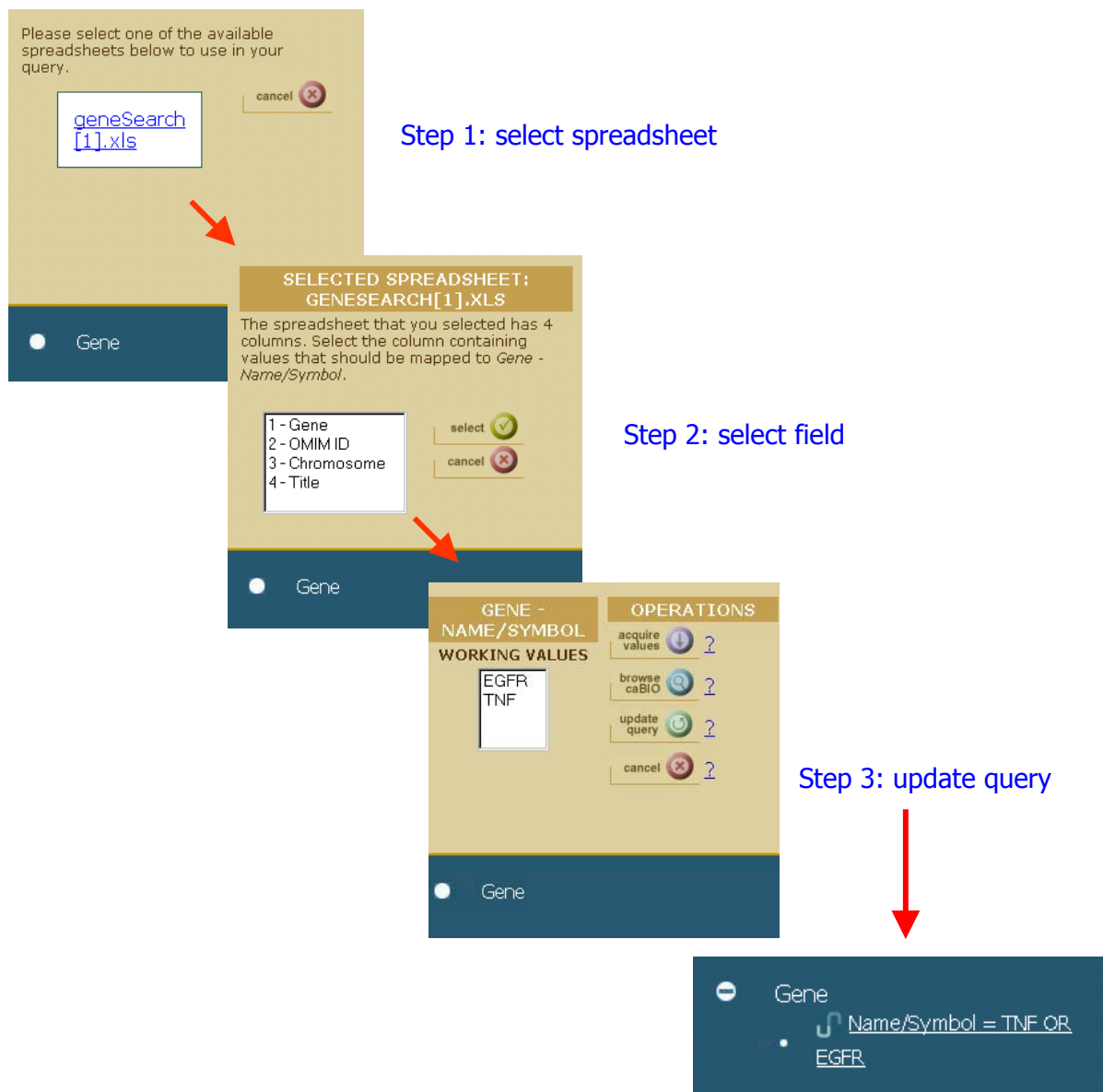
The file location can be entered using the **Browse** button, and the worksheet number should be left as the default, 1. Because our sample spreadsheet includes a header in row number 1, we also specify row 1 for option C, and row 2 for the starting row of the actual data. After setting these values, click the **Select** button to upload the spreadsheet file to the BIOgopher server.

After the spreadsheet has loaded, the right-hand display will include the new filename in the list of CURRENTLY AVAILABLE SPREADSHEETS. Loading a spreadsheet does not in itself affect any queries, and any number of spreadsheet files can be uploaded by repeating the above steps. The loaded spreadsheets are simply made available for acquiring values and for merging previous data with new search results.

This example will use the gene names in the saved spreadsheet to formulate a new query. To start building the query, select the search criteria tab in the left panel, and select the Gene object from the pull-down menu in slot A. This time we will be searching for pathways associated with genes, so we name the new query “PathQuery,” and click on the **Create** button. This should take you to a screen identical to Figure 4.2-2.

As in the previous example, we begin specifying our search criteria by selecting the Gene object’s **Name/Symbol** attribute. In this case however, instead of typing in the gene names manually, we will acquire them from our spreadsheet using the **acquire values** button. Figure 4.3-2 summarizes the sequence of screens that enable BIOgopher’s values acquisition.

Each of the first three snapshots in Figure 4.3-2 corresponds to what the user sees in the right panel as she steps through the values acquisition process. In step 1, the user selects a spreadsheet from the list of all loaded spreadsheets. Step 2 requires you to select the field in the loaded spreadsheet that corresponds to the currently selected attribute. In this case we are working with the gene’s Name/Symbol, and we select the first field, “Gene.” The third snapshot shows the values that were loaded from the spreadsheet for this field. BIOgopher allows you to further edit these values before updating the query, but for this example we simply accept them without any further editing. Finally, clicking on **update query** in step 3 modifies the lower right-hand query panel to reflect these new values.



**Figure 4.3-2 Acquiring Values From a Spreadsheet**

The query now looks more or less similar to that which we began building in Example 1. A new “S”-shaped symbol however, now appears in front of the **Name/Symbol** attribute. This symbol is used to indicate that the new results obtained in this search will *not* be merged with the fields currently defined in the existing spreadsheet. Clicking on this icon will force the results to be merged—a capability we will explore in Example 3.

As in the previous example, we again constrain the gene search to human genes, this time by selecting the Taxon object’s **Abbreviation** attribute, and using the **add value** button to enter “Hs”. This completes our search criteria specifications, and we can now initialize the report format by clicking on the **report format** tab in the left panel. Naming the report “PathReport” and clicking on the **Create** button generates the screen in Figure 4.3-3.

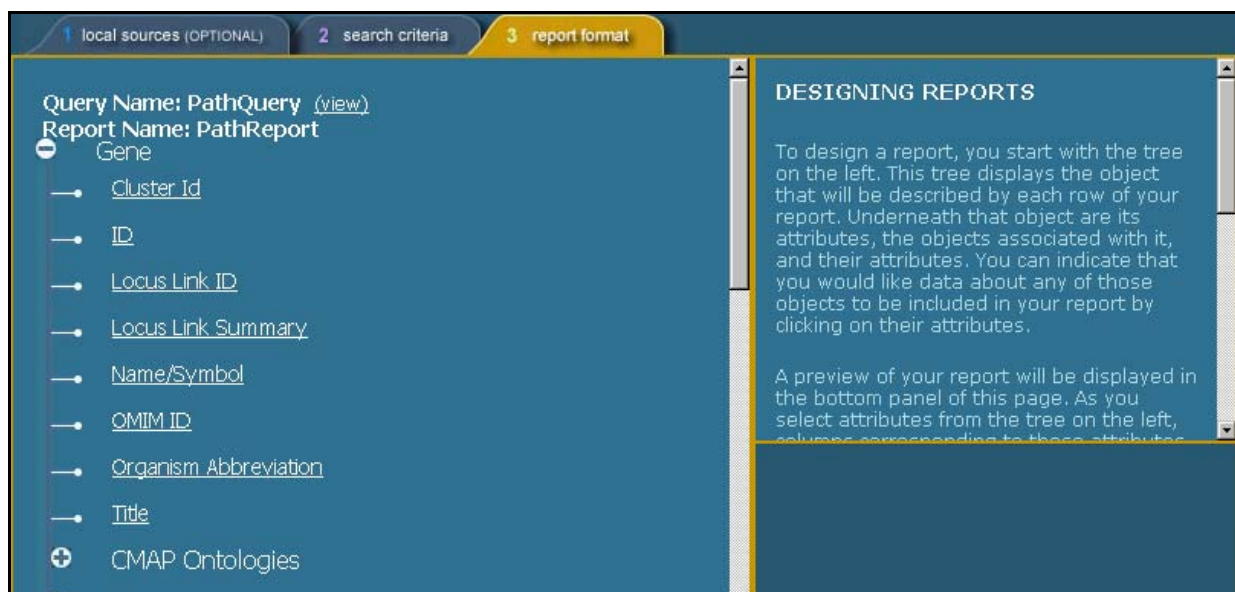


Figure 4.3-3 Selecting Fields for the Report

The first attribute we select is again the gene's **Name/Symbol** attribute. In this example we are also interested in learning about the pathways associated with the identified genes, so we scroll down to the Pathway object and select two attributes, **Name** and **Display Value**. Figure 4.3-4 shows the resulting report format.

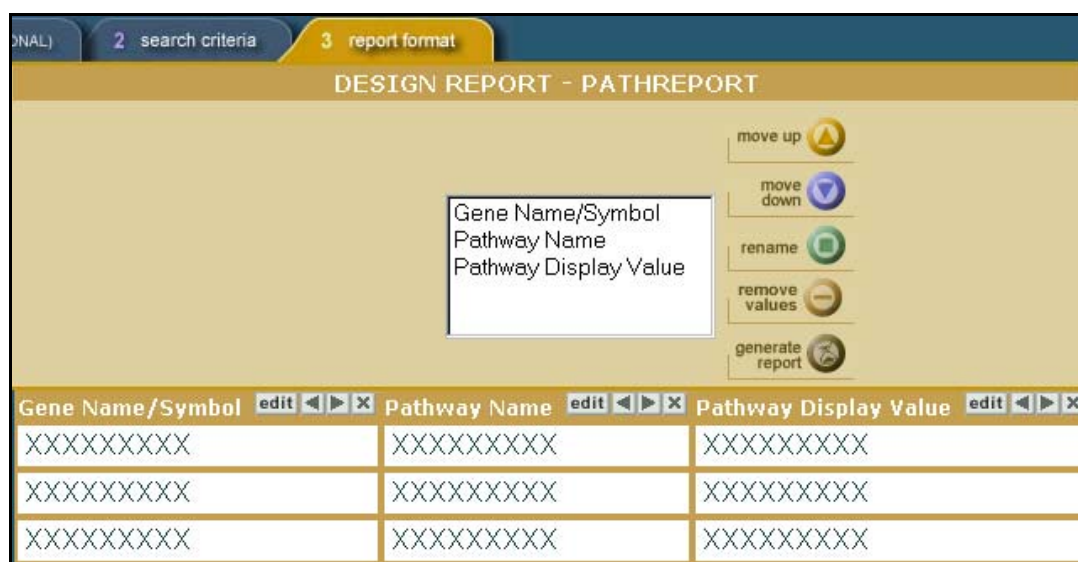
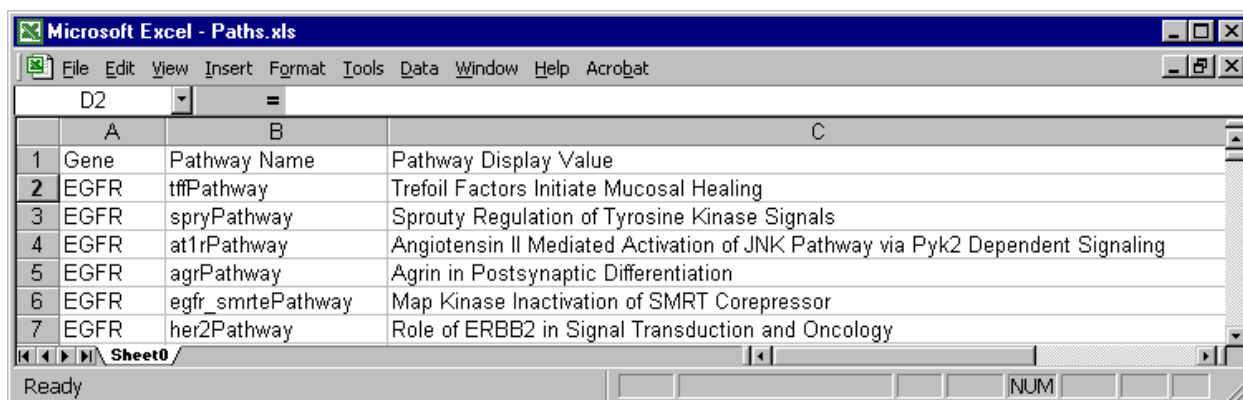


Figure 4.3-4 Report Design for Pathway Information

Again, we simplify the “Gene Name/Symbol” field to just “Gene,” and press the **generate report** button. Downloading the results table to an excel file then produces a spreadsheet whose first several rows are shown in Figure 4.3-5.



	A	B	C
1	Gene	Pathway Name	Pathway Display Value
2	EGFR	tffPathway	Trefoil Factors Initiate Mucosal Healing
3	EGFR	spryPathway	Sprouty Regulation of Tyrosine Kinase Signals
4	EGFR	at1rPathway	Angiotensin II Mediated Activation of JNK Pathway via Pyk2 Dependent Signaling
5	EGFR	agrPathway	Agrin in Postsynaptic Differentiation
6	EGFR	egfr_smrtePathway	Map Kinase Inactivation of SMRT Corepressor
7	EGFR	her2Pathway	Role of ERBB2 in Signal Transduction and Oncology

Figure 4.3-5 Results Downloaded to an Excel Spreadsheet

#### 4.4 Example 3: Merging New Results With A Local Spreadsheet

In Example 2 we loaded *geneSearch.xls* (from Example 1) onto BIOgopher and used this spreadsheet to acquire values for the Gene object's **Name/Symbol** attribute. In this last example we also acquire values from *geneSearch.xls*, but this time we will selectively merge the new results with those in the spreadsheet.

We begin by loading the spreadsheet as outlined in [Section 4.3](#), and continue by acquiring the values for the Gene object's **Name/Symbol** attribute as described in Figure 4.3-2. In this case however, we click on the “S”-shaped icon in the query panel to indicate that the new results should be *merged* with those in the spreadsheet. Finally, we again add the additional constraint that the taxon abbreviation is “Hs”. Figure 4.4-1 shows the new query.

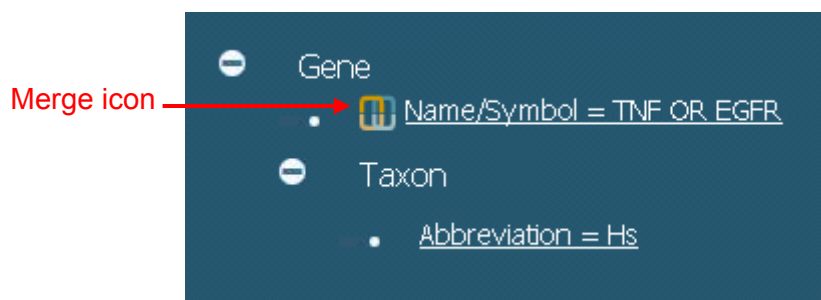


Figure 4.4-1 Using the Merge Icon to Combine Results

The new results to be merged with those previously obtained can now be defined using the **report format** tab. Clicking on that tab brings up the form that was shown in Figure 4.2-7 for initializing a report format. Give the report a new name such as “mergeReport.” Pressing the **Create** button then takes you to report design form shown in Figure 4.4-2.

The report is already initialized with the columns appearing in the spreadsheet. The additional columns that we can now select from the Gene object's attributes and related objects will define the new results obtained in this search. We begin however, by removing the “Title” field from the current report format, as these tend to be relatively large text fields. To do this, select “Title,” as indicated in Figure 4.4-2, and click the remove values button.

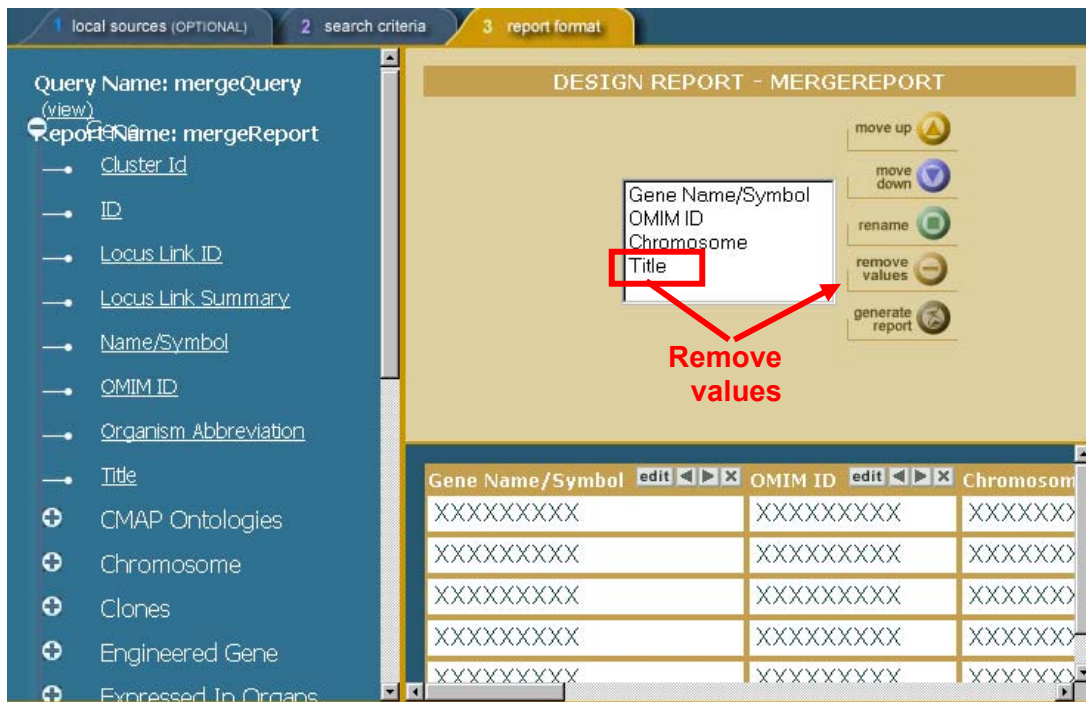


Figure 4.4-2 Removing Fields From the Previous Results

We complete this exercise by adding one new field to the report format, the MapLocation object's **Location** attribute. Figure 4.4-3 shows the final report format and the resulting output.

The screenshot shows the 'DESIGN REPORT - MERGEREPORT' interface. The main area displays a list of fields: Gene Name/Symbol, OMIM ID, Chromosome, and Map Location. A red box highlights the 'Map Location' field, and a red arrow points to the 'remove values' button. The 'remove values' button is located in the top right corner of the main area, along with other buttons like 'move up', 'move down', 'rename', and 'generate report'.

Gene Name/Symbol	OMIM ID	Chromosome	Map Location
XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX
XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX
XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX	XXXXXXXXXX

	A	B	C	D
1	Gene Name/Symbol	OMIM ID	Chromosome	Map Location
2	EGFR	131550	7	7p12
3	TNF	191160	6	6p21.3

Figure 4.4-3 The Final Report Format and Resulting Output



Two things should be noted in this example. First, because the *Gene Name/Symbol* field must be associated with a specified column of a local spreadsheet, it is *not* possible to rename this column.

Second, it is important to understand from which sources the reported values are derived. The values in those columns that were included in the original spreadsheet are *not* regenerated when the new report is created. These previously obtained values are simply merged with the new results. Only those fields that are explicitly created as *new* columns in the report will contain values obtained from the new search.

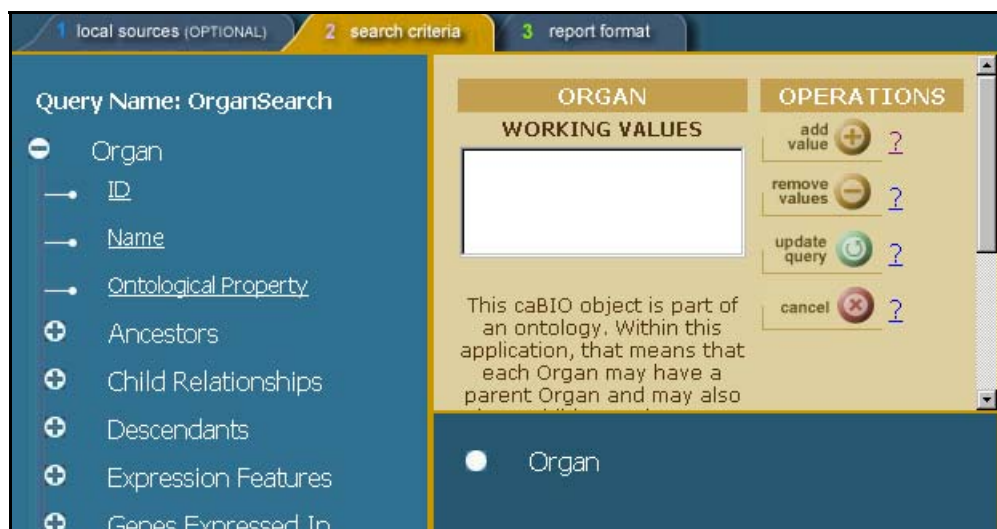
Thus, if the new search imposes additional constraints that were not included in the spreadsheet query, it is possible that the merged report may contain logical inconsistencies. As a simple example, consider what might happen if the taxon constraint was not imposed when the spreadsheet was generated, but the new results were obtained under the constraint that “Taxon Scientific Name = Homo sapiens.” The chromosome names in the spreadsheet may refer to any species, but the map locations associated with the gene names/symbols will be associated with Homo sapiens only. Finally, this could lead to results such as those shown in Table 4.4-1.

**Table 4.4-1 Erroneous Results Obtained From Inconsistently Merged Values**

<u>Gene Name/Symbol</u>	<u>Chromosome</u>	<u>Map Location</u>
TGFB1	7	19q13.1
MYC	20	8q24.12-q24.13

## 4.5 Accessing the EVS Terminologies from BIOgopher

In most cases, selecting an attribute and clicking the **add value** button will generate a textbox where the user can type in a text string. For attributes named **Ontological Property** however, this action will instead generate an interface to the EVS server. Consider for example, the object tree under the Organ object shown in figure 4.5-1.



**Figure 4.5-1 The Organ Object Tree**

Selecting the **Ontological Property** attribute and clicking the **add value** button will in this case generate the EVS screen shown in Figure 4.5-2. There are two ways of interacting with this

screen. First, you can simply enter a search term in the upper left-hand panel and press the **Search** key. The results will appear in the lower left-hand panel, in the area titled “Search Results.”

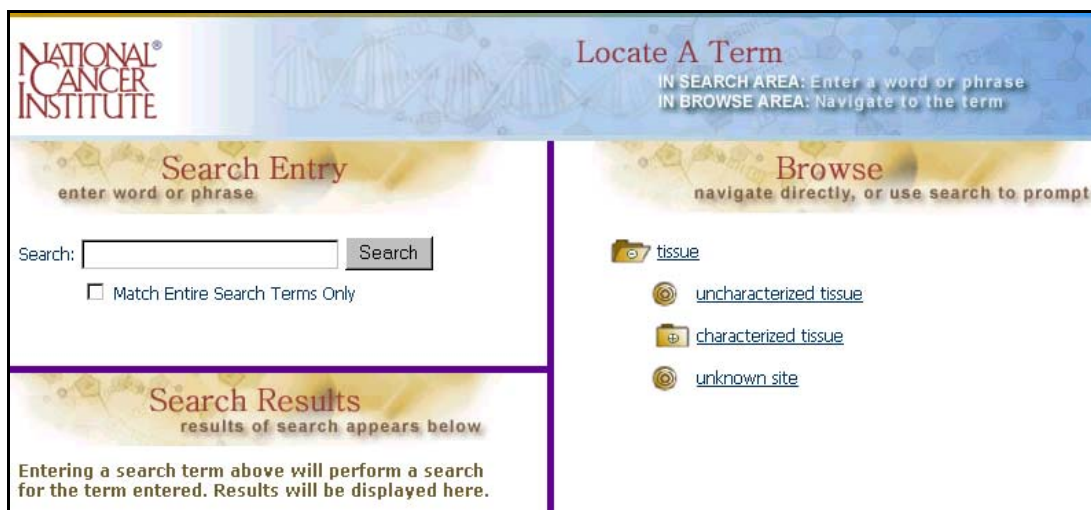


Figure 4.5-2 The EVS Search Interface

Figure 4.5-3 illustrates how the display changes after entering “heart” in the search textbox, checking **Match Entire Search Terms Only**, and pressing the **Search** button. Each matching term in the Search Results area is displayed as hyperlinked text and has a button labeled **OPEN TO BROWSE** associated with it. Clicking on this button will expose that term in the navigation tree in the right-hand panel, with the selected term highlighted and appearing at the top of the tree.

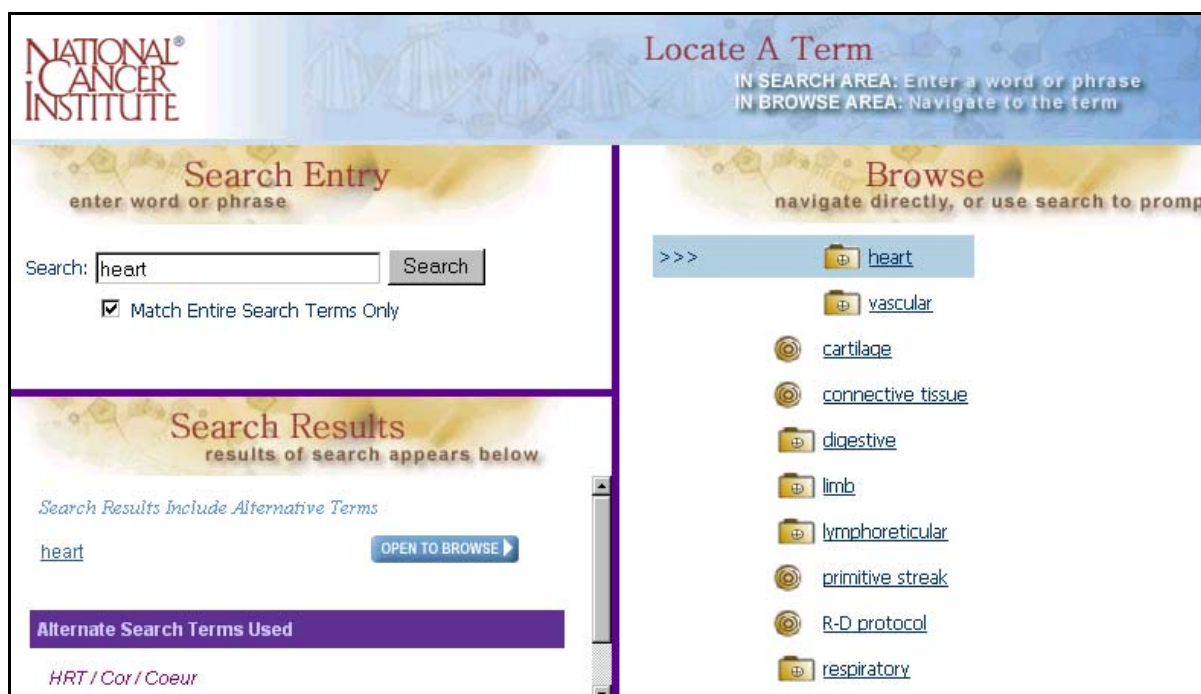


Figure 4.5-3 Using the Search Area to Locate Terms

Alternatively, clicking on the hyperlinked text itself will “select” that term as the value you wish to insert as a working value in the BIOgopher interface. Before inserting the term however, a pop-up dialog box will prompt the user for confirmation.

The second way to interact with the EVS interface is to browse the navigation tree directly. Clicking on a hyperlinked term in the tree has the same effect as clicking on a term in the Search Results area—a pop-up dialog will prompt the user for confirmation of that term as a working value. Figure 4.5-4 summarizes the icons used to expand and collapse the branches of the tree.

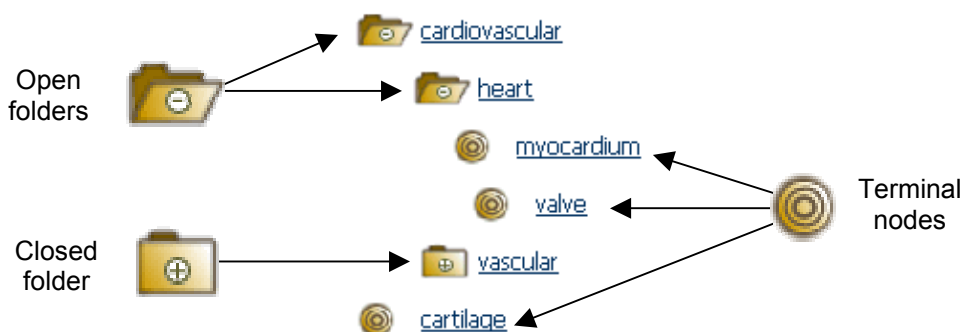


Figure 4.5-4 Icons for browsing the EVS navigation tree

## 4.6 The caBIO Data Sources

The caBIO domain objects were developed primarily in response to the need for programmatic access to the information at several NCI web sites, including:

- the Cancer Genome Anatomy Project ([CGAP](#))
- the CGAP Genetic Annotation Initiative ([GAI](#))
- the Enterprise Vocabulary Services ([EVS](#))
- the Cancer Data Standards Repository ([caDSR](#))
- the Mouse Models of Human Cancers Consortium ([MMHCC](#))
- the Cancer Molecular Analysis Project ([CMAP](#))
- the Gene Expression Data Portal ([GEDP](#))

While much of this information is in theory available from multiple public sites, the number of links to traverse and the extent of collation that would have to be performed is daunting. The CGAP, CMAP, and GAI web sites have distilled this information from both internal and public databases, and the caBIO data warehouses have optimized it for access with respect to the types of queries defined in the APIs. In this section we discuss the external and internal data sources for caBIO and how the information these sources provide can be accessed via caBIO objects.

Although the caBIO data are extracted from many sources that include information from a wide variety of species, we emphasize that *only genomic data pertaining to human and mouse are available from caBIO*. caBIO provides access to curated data from multiple sources, including:

- The NCBI [UniGene](#) database [1]. Unigene provides a non-redundant partitioning of the genetic sequences contained in GenBank into gene clusters. Each such cluster has a unique UniGene ID and a list of the mRNA and expressed sequence tag (EST) sequences that are subsumed by that cluster. Related information stored with the cluster includes tissue types in



which the gene has been expressed, mapping information, and the associated LocusLink, OMIM, and HomoloGene IDs, thus providing access to related information in those NCBI databases as well.

- NCBI's [LocusLink](#) database [2]. LocusLink contains curated sequence and descriptive information associated with a gene. Each entry includes information about the gene's nomenclature, aliases, sequence accession numbers, phenotypes, UniGene cluster IDs, OMIM IDs, gene homologies, associated diseases, map locations, and a list of related terms in the Gene Ontology Consortium's ontology. Sequence accessions include a subset of GenBank accessions for a locus, as well as the NCBI Reference Sequence. As mentioned above, a caBIO Gene object has explicit methods for retrieving the gene's associated LocusLink, OMIM, and Unigene IDs.
- The [Gene Ontology Consortium](#) [3]. The Gene Ontology Consortium provides a controlled vocabulary for the description of molecular functions, biological processes, and cellular components of gene products. The terms provided by the consortium define the recognized attributes of gene products and facilitate uniform queries across collaborating databases. caBIO does not extract ontology terms directly from the Gene Ontology Consortium but, instead, extracts those terms stored with the LocusLink entry for that gene.
- The [HomoloGene](#) database [4]. HomoloGene is an NCBI resource for curated and calculated gene homologs. The caBIO data sources capture only the calculated homologs stored by HomoloGene. These calculated homologs are the result of nucleotide sequence comparisons performed between each pair of organisms represented in UniGene clusters.
- [BioCarta](#) pathways. BioCarta and its Proteomic Pathway Project (P3) provides detailed graphical renderings of pathway information concerning adhesion, apoptosis, cell activation, cell signaling, cell cycle regulation, cytokines/chemokines, developmental biology, hematopoiesis, immunology, metabolism, and neuroscience. NCI's CMAP web site captures pathway information from BioCarta, and transforms the downloaded image data into Scalable Vector Graphics ([SVG](#)) representations that support interactive manipulation of the online images. The CMAP web site displays BioCarta pathways selected by the user and provides options for highlighting *anomalies*, which include under- or over-expressed genes as well as mutations. The caBIO Pathway objects make this same information available through BIOgopher.
- The Cancer Genome Anatomy Project [5]. The NCI CGAP web site provides a collection of gene expression profiles of normal, pre-cancer, and cancer cells taken from various tissues. The CGAP interface allows the user to browse these profiles by various search criteria, including histology type, tissue type, library protocol, and sample preparation methods. The goal at NCI is to exploit such expression profile information for the advancement of improved detection, diagnosis, and treatment for the cancer patient. Researchers have access to all CGAP data and biological resources for human and mouse, including ESTs, gene expression patterns, SNPs, cluster assemblies, and cytogenetic information via the BIOgopher interface.
- The CGAP Genetic Annotation Initiative [6]. GAI is an NCI research program to explore and apply technology for identification and characterization of genetic variation in genes important in cancer. The GAI utilizes data-mining to identify "candidate" variation sites from

publicly available DNA sequences, as well as laboratory methods to search for variations in cancer-related genes. All GAI candidate, validated, and confirmed genetic variants are available directly from the GAI web site, and all validated SNPs have been submitted to the NCBI dbSNP database as well. SNPs identified by the GAI project can be accessed using the BIOgopher *SNP* objects. The sequencing trace files used by GAI are imported from [Washington University](#).

- The NCI Cancer Therapy Evaluation Program [7]. CTEP funds an extensive national program of basic and clinical research to evaluate new anti-cancer agents, with a particular emphasis on translational research to elucidate molecular targets and drug mechanisms. In response to this emergent need for translational research, there has been a groundswell of translational support tools defining controlled vocabularies and registered terminologies so as to enhance electronic data exchange in areas that have heretofore been relatively non-computational. The caBIO trials data are updated with new CTEP data on a quarterly basis.
- NCI's Cancer Molecular Analysis Project [8]. The CMAP web site is implemented using the caBIO domain objects available through the caCORE API described in the caCORE Technical Guide. The goal of CMAP is to enable researchers to identify and evaluate molecular targets in cancer. Towards this goal, CMAP provides four interfaces. The CMAP *Profile Query* tool finds genes with the highest or lowest expression levels (using SAGE and microarray data) for a given tissue and histology. Selecting a gene from the resulting table leads to a *Gene Info* page, providing information about cytogenetic location, chromosome aberrations, protein similarities, curated and computed orthologs, and sequence-verified as well as full-length MGC (Mammalian Gene Collection) clones, along with links to various other databases.

CMAP's *Molecular Targets* interface organizes collections of genes by pathways and by ontology. Two ontologies are available: (1) the GO ontology described above, and (2) the CMAP ontology described here. The CMAP ontology relates functional classifications to molecular targets and agents. CMAP's *AgentSearch* tool allows the researcher to search for drug therapies by name (with wildcard matching), with the option of restricting the search to agents that are either associated with a term in the CMAPOntology or registered with a CTEP protocol. If the agent is associated with CTEP protocols, a table is presented on the *Agent Info* page, listing the title of each protocol and a link to its associated documentation. Selecting an entry from this table in turn leads to the *Therapeutic Trials Info* page for that CTEP protocol. With the exception of the Mitelman Chromosome Aberration data, all of the information available through CGAP is also accessible through the BIOgopher objects.

- The NCI Enterprise Vocabulary Services [9] ([EVS](#)). The EVS provides NCI with services and resources for controlled biomedical vocabularies, and includes both the NCI Thesaurus and the NCI Metathesaurus. The Thesaurus is composed of over 27,000 concepts represented by about 78,000 terms. The Thesaurus is organized into 18 hierarchical trees covering areas such as Neoplasms, Drugs, Anatomy, Genes, Proteins, and Techniques. These terms are deployed by NCI in its automated systems for uses such as keywording and database coding. The NCI Metathesaurus maps terms from one standard vocabulary to another, facilitating collaboration, data sharing, and data pooling for clinical trials and scientific databases. The

Metathesaurus is based on the NLM's Unified Medical Language System and is composed of over 70 biomedical vocabularies.<sup>19</sup>

## References

1. Schuler (1997). Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J Mol Med* 75(10):694–8.
2. Pruitt KD, and Maglott DR (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29(1):137–40.
3. The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25–9.
4. Zhang, Schwartz, Wagner, and Miller (2000). A Greedy algorithm for aligning DNA sequences, *J Comp Biol* 7(1-2):203–14.
5. Strausberg RL (2001). The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. *J Pathol*, 195:31–40.
6. Clifford R, Edmonson M, Hu Y, Nguyen C, Scherpbier T, and Buetow KH (2000). Expression-based genetic/physical maps of single-nucleotide polymorphisms identified by the cancer genome anatomy project. *Genome Res* 10(8):1259–65.
7. Ansher SS, Scharf R (2001). The Cancer Therapy Evaluation Program (CTEP) at the National Cancer Institute: industry collaborations in new agent development. *Ann N Y Acad Sci.* 949:333-40.
8. Buetow KH, Klausner RD, Fine H, Kaplan R, Singer DS, Strausberg RL (2002). Cancer Molecular Analysis Project: Weaving a rich cancer research tapestry. *Cancer Cell* 1(4):315-8.
9. Hartel, F.W. and de Coronado, S (2002). Information Standards within NCI. In: *Cancer Informatics: Essential Technologies for Clinical Trials*. J. S. Silva, M. J. Ball, C. G. Chute, J. V. Douglas, C. Langlotz, J. Niland and W Scherlis, eds. Springer-Verlag.

---

<sup>19</sup> See [Chapter 2](#) of this manual for a more in-depth discussion of the EVS

## 5.0 THE caARRAY PROJECT

The recently completed draft human genome sequence indicates that the human genome comprises only about 30,000 genes—just twice the number as the fly or worm. Yet with these 30,000 genes the human genome generates approximately 90,000 proteins, due to alternative splicing of the original nucleotide sequences. Only a fraction of these genes and their products are currently associated with known function however, and the further elucidation of their roles in disease processes has become a central focus in today's research.

With the advent of DNA microarray technology, it is now possible to monitor the expression of almost every gene in a given genome on a single chip, and a typical microarray experiment will yield millions of data points. The massive amount of microarray data being generated today presents a significant challenge for analysis, storage, and exchange of data. The databases and tools provided by the caArray project at NCI and described in this chapter were developed to address these issues.

The NCI Gene Expression Data Portal provides a repository for secure storage of researchers' microarray data and facilitates the exchange of pre- and post-publication data. The GEDP hosts a public web site providing browse/download capabilities and data analysis to all users, along with data submission tools for registered users. The GEDP supports the [MIAME](#) standard, and the GEDP object model is based on the MicroArrayGeneExpression Object Model ([MAGE-OM](#)).

The analysis tools provided with GEDP allow scientists to analyze private as well as publicly available microarray data online. The [XpressionWay](#) and [Affy Cel File Analysis Center](#) tools are fully integrated with the GEDP database and can be invoked from the GEDP interface. XpressionWay allows users to compare the expression levels of two experiments in the context of molecular pathways. Using diagrams provided by [BioCarta](#), XpressionWay allows the user to selectively highlight those genes that are comparatively down or up-regulated. The Affy Cel File Analysis Center allows users to search Affymetrix cel files for values pertinent to specific genes; to browse cel files interactively; and to download NetCDF versions of cel files for use with external statistical packages.

Other tools provided by the caArray project include the stand-alone [caWorkbench](#) and the [webCGH](#) tools. caWorkbench is an extendible and flexible desktop tool for microarray data analysis and visualization. The current release (1.0) includes several data analysis and visualization functions, including hierarchical clustering, self-organizing maps, color mosaic images, and biological pathway visualization.

webCGH is a web-based application for visualizing and mining microarray-based Comparative Genomic Hybridization data. webCGH enables users to search for CGH experiments in a database; to create persistent user-defined groupings of experimental bioassays; and to generate whole genome plots with zoom capabilities for focusing on chromosomal regions of interest. In webCGH *line* plots, DNA copy measurements are plotted relative to genome position; in webCGH *annotation* plots DNA copy measurements are shown relative to annotated genome features.

This chapter begins with step-by-step instructions on using the GEDP interface for browsing, downloading, submitting, and analyzing microarray data. Section 5.2 then describes the desktop caWorkbench tool, and Section 5.3 concludes the chapter with a discussion of the Comparative Genome Hybridization visualization tool, webCGH.

## 5.1 The Gene Expression Data Portal

The object model that defines the GEDP database at NCI is derived from the MAGE-OM. Like MAGE-OM, the GEDP object model supports all DNA arrays, including spotted and synthesized arrays, and oligo-nucleotide and cDNA arrays—independent of image analysis and/or data normalization methods. Because the model also allows for comprehensive annotation of experimental results, the GEDP serves as a repository for both raw and processed data.

To access GEDP, begin by pointing your browser to <http://gedp.nci.nih.gov>. The GEDP home page will appear, as shown in Figure 5.1-1. First-time visitors are encouraged to visit the link at the top of the page, which allows users to register new accounts.

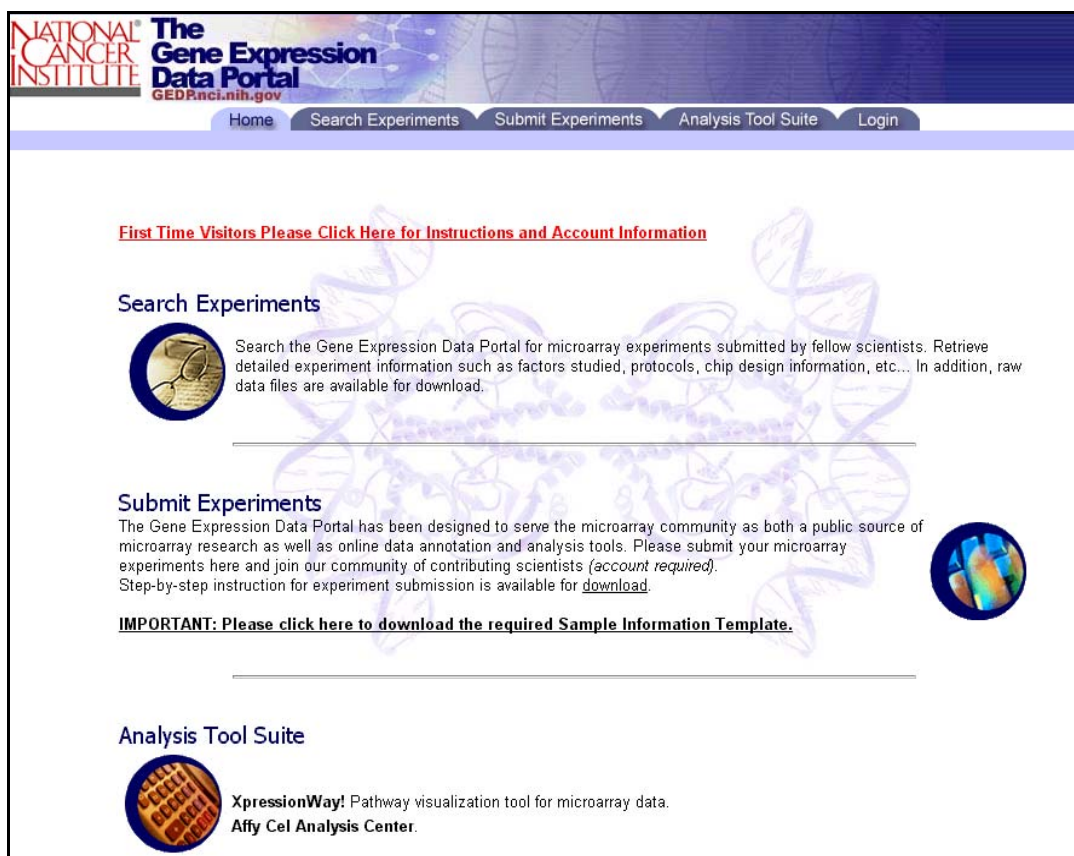


Figure 5.1-1 The GEDP Welcome Page

Folder tabs running across the top of all of the GEDP screens provide access to the *Home* page (Figure 5.1-1), the *Search Experiments* pages, the *Submit Experiments* pages, the *Analysis Tool Suite*, and the *Login* screen. All data submissions use standardized submission forms; the home page also provides a convenient shortcut for downloading the required Sample-Specific Information Template. The three main pages for search, submission, and analysis can also be accessed by clicking on the associated graphical icons on the home page.

### 5.1.1 GEDP's Search Tools

Figure 5.1-2 shows the Basic Search form for exploring the GEDP. Users can search for experiments based on any combination of experiment IDs, investigator names, species, tissue



names, platform types, and image analysis software. Only the **Experiment ID** field allows text entry; the remaining search criteria provide picklists and controlled vocabularies for selection.

**Figure 5.1-2 GEDP’s Basic Search Form**

The default values shown in Figure 5.1-2 are all-inclusive; pressing on the **Search** key with these settings will retrieve all publicly available microarray data stored in GEDP. Those fields that are left blank—**Experiment ID** and **Tissue**—imply that all experiment IDs and all tissues are acceptable. Clicking the **Reset** key restores all fields on the search form to their default values.

Entries in the **Investigator’s Name** field are chosen from a list of registered GEDP Principal Investigators (PIs). Choices for the **Species** field include human, mouse, rat, Drosophila, yeast, and *other*. In order to select an entry for **Tissue Name**, you must first select a species. Pressing the **Select** key next to the tissue field will then open a pop-up window to an appropriate branch of the EVS tissue taxonomy for that species.<sup>20</sup> Figure 5.1-3 shows the EVS Navigator window after the user has selected “human” as the species and pressed the **Select** key. In this example, the user has entered the search term “breast” in the SEARCH: Entry area and selected **SHOW in DATATREE** for that term in the SEARCH: Results pane.

The EVS window is divided into a Search panel (left) and a DataTree panel (right). The top region of the Search panel (SEARCH: Entry) provides a textbox for entering search terms. The EVS performs semantic as well as orthographic matching, and the search results shown in the lower half of the Search panel (SEARCH: Results) will include syntactic matches as well as alternative

<sup>20</sup> The Enterprise Vocabulary Services provide controlled vocabulary terminologies for many of the applications at NCI, as described elsewhere in this manual, as well as in the NCICB Technical Guide.

terms or aliases for the search terms. Each match in the SEARCH: Results area has an icon labeled **SHOW in DATATREE**. Selecting this icon causes that branch of the DataTree to be highlighted, as shown in Figure 5.1-3.

The screenshot shows the EVS Navigator Window with two main panels. The top header includes the National Cancer Institute logo and EVS Enterprise Vocabulary Services. Below the header, a purple bar contains instructions: "TO LOCATE A TERM: 1 Enter a word or phrase in SEARCH /or/ 2 Navigate to the term in the DATA TREE".

The left panel is titled "SEARCH: Entry" and "Enter Search Word or Phrase Below". It contains a search input field with the text "breast" and a "SEARCH" button. Below the input field is a checkbox labeled "Match Entire Search Terms Only".

The right panel is titled "DATA TREE" and "Browse DataTree Directly, Or Use Search for Prompting". It displays a hierarchical list of body systems: Digestive System, Endocrine System, Exocrine System, Genitourinary System, Hematopoietic Body System, Immune System, and Integumentary System. The "Integumentary System" is expanded, showing a sub-item "Breast" which is highlighted with a blue background. Below the "Breast" item are "Skin" and "Musculoskeletal System".

Below the "SEARCH: Entry" panel is the "SEARCH: Results" section, titled "Results of Your Search Will Appear Below". It includes a note "Search Results Include Alternative Terms" and a list of search results: "Breast", "Mammary Epithelium", "Mammary Gland", and "Mammary Gland Fat-pad". Each result has a "SHOW in DATATREE" button to its right.

**Figure 5.1-3 The EVS Navigator Window**

Clicking on a term in the DataTree panel will cause that item to be selected as the candidate entry for the tissue field in the GEDP Basic Search form. A pop-up dialog box will appear on the screen to confirm this choice, and pressing **OK** will close the EVS window and insert that value in the GEDP search form.

Pressing **Cancel** allows the user to continue browsing the tree for more specific terms. All terms in the DataTree having more specific “sub-terms” associated with them have folder-like icons to the left of the term. Pressing on those terms whose folder icons are labeled with a “+” sign expands that term, as in the example above, where “Integumentary System” has been expanded. Pressing on an expanded folder icon will “collapse” that term and return the tree to its previous state.

The remaining fields on the GEDP’s Basic Search form are populated using pull-down lists that provide choices for the sample type, platform type, and analysis software that may have been applied to the data. Pressing on the **Search** key after filling out the search criteria fields initializes the search, and the results are then presented on a Search Results page, as shown in Figure 5.1-4.

This initial results page is but a summary of the results, and shows limited information about the experiment(s) retrieved. Selecting the **View** button in the first column for the experiment you are interested in will bring up the detail on that experiment.

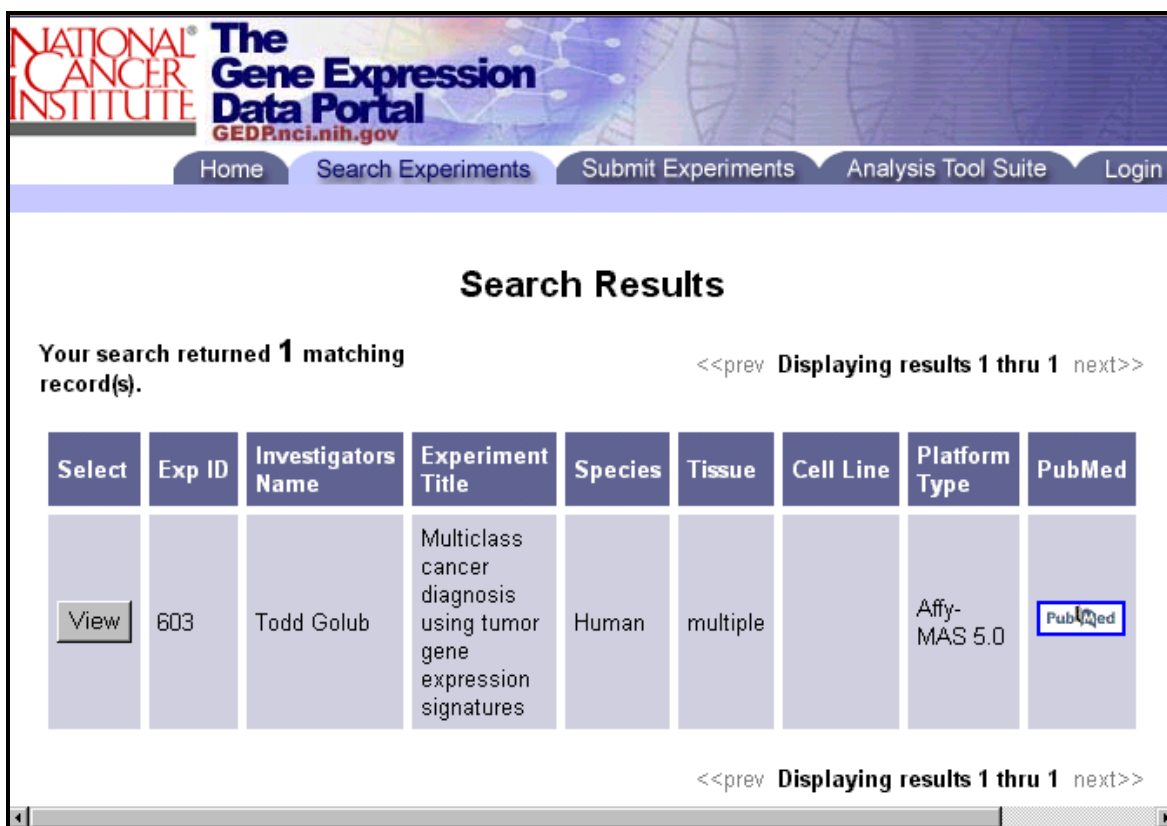


Figure 5.1-4 The GEDP Search Results Page

Figure 5.1-5 shows an Experiment Detail page. Many—though not all—of the fields on the details page correspond to the search criteria fields on the Basic Search form. If the experiment has multiple species, tissues, or cell lines, the list of values will be enclosed in the “[ ]” corresponding to the field.

Data files can be downloaded from this Experiment Detail page by selecting the **Download Data Files** radio button and pressing the **Submit** key. Selecting the download option causes the Experiment Detail page to be refreshed, with the bottom of the page now presenting an array of icons for downloading the various files associated with that experiment.

For example, Affymetrix experiments may have associated .txt files, .cel files, and .dat files. For GenePix experiments, icons to download the .gal and .gpr files will appear. There may also be icons for downloading the MAGEML document generated upon experiment submission as well as the sample information file (sample.txt) submitted with all experiments. As shown in Figure 5.1-5, the Experiment Detail page also provides access to any PubMed abstracts associated with the experiment.

The second radio button on the Experiment Detail page, **Retrieve and Analyze Data Sets with Array Design**, initializes the XpressionWay pathway visualization tool, which is covered in [Section 5.1.3](#).





**The Gene Expression Data Portal**  
 GEDP.nci.nih.gov

[Home](#)
[Search Experiments](#)
[Submit Experiments](#)
[Analysis Tool Suite](#)
[Login](#)

### Detailed Experiment Information

<b>Investigator</b>	Donna Albertson	<b>Experiment Contact</b>	Donna Albertson Comprehensive Cancer Center, University of California San Francisco
<b>Experiment Type</b>	Normal vs. Diseased Comparison	<b>Experiment Title</b>	Assembly of microarrays for genome-wide measurement of DNA copy number
<b>Species</b>	[Human]	<b>Chip Name</b>	[HumArray1.14]
<b>Tissue</b>	[ ]	<b>Cell Line</b>	[Cell line MPE600, Cell line HCT116, Cell line GM03134, Cell line GM13031, Cell line MDA-MB-231, Cell line GM03563, Cell line T47D, Cell line S0034, Cell line GM01218, Cell line GM03576, Cell line GM02948, Cell line GM10315, Cell line SW837, Cell line GMGM07408, Cell line S1514, Cell line HT29, Cell line GM04435, Cell line GM01750, Cell line GM13330, Cell line GM07081, Cell line GM05296, Cell line GMM01524, Cell line MDA-MB-453, Cell line GM00143, Cell line COLO320]
<b>Scanner Hardware</b>		<b>Software</b>	[Image Analysis Software : UCSF SPOT]
<b>Experiment Description</b>	Assembled arrays of approximately 2400 BAC clones for measurement of DNA copy number across the human genome. This study is published in Nat. Genet. 29(3):263-4, 2001		


[View PubMed Abstract](#)

☒ Download Data Files
   
☐ Retrieve and Analyze Data Sets with Array Design: HumArray1.14 ▾

**Figure 5.1-5 The Experiment Detail Page**

## 5.1.2 Submitting Experimental Data

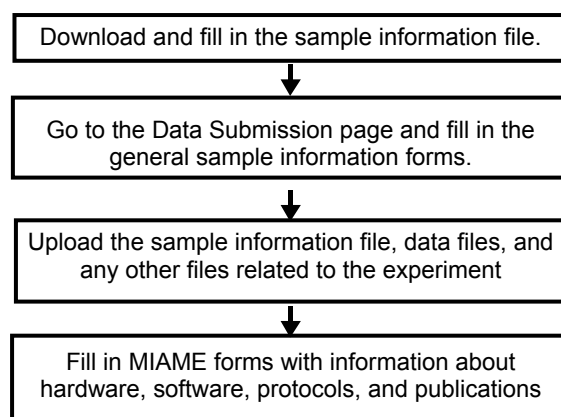
The GEDP interface makes the submission process as easy as possible while preserving the rich set of annotations specified in the MIAME standard. The GEDP provides explicit support for several Affymetrix (MAS4, MAS5, and Ann Arbor Suite) and spotted array (GenePix and UCSF SPOT) platforms. While not explicitly supported, experimental data obtained using other platforms can also be submitted by selecting the “other” platform type. Any data files submitted for these other platforms are available to other users for download, but are not amenable to analysis by the tools provided with the GEDP.

While browsing and downloading can be performed anonymously, submission of experimental data requires a registered user account. Several files are required with all data submissions; the specific files that are required will depend upon the experimental platform. Table 5.1-1 lists the platform-specific files that are required and optional.

**Table 5.1-1 Platform-Specific Files for Experiment Submission**

<i>Platform</i>	<i>File Name</i>	<i>File Extension</i>	<i>Status</i>
<b>Affymetrix Chips</b> MAS4, MAS5, Ann Arbor	Cell Intensity File	*.cel	<b>required</b>
	Data File	*.txt	optional
	Report File	*.rpt	optional
	Image Data File	*.dat	optional
	Chip File	*.chp	optional
	Sample Info File	sample.txt	<b>required</b>
<b>Spotted Arrays</b> GenePix 3.0	GenePix Array List File	*.gal	<b>required</b>
	GenePix Results File	*.gpr	<b>required</b>
	Sample Info File	sample.txt	<b>required</b>
UCSF SPOT	Chip Clone File	spotclone.txt	<b>required</b>
	Summarized Result File	sproc.txt	<b>required</b>
	Numeric Raw Data File	spot.txt	<b>required</b>
	Spot File	*.spt	optional
	BAC Clone File	clonepos.txt	optional
	Raw Image Data File	*.tif	optional
	Sample Info File	sample.txt	<b>required</b>
<b>Other</b>	Any file	*.*	optional

Clicking on the *Submit Experiments* tab at the top of any GEDP screen initiates the submission process. This will bring up a Login page for submitting data, which will provide an opportunity for creating an account if you have not already done so. Figure 5.1-6 outlines the general flow of the data submission process.



**Figure 5.1-6 Experiment Submission Process**

**Step 1.** The first step of the submission process involves downloading the Sample-Specific Information Template from the GEDP home page using the link at the bottom of that screen. The download dialog that appears provides several templates for different species and experimental

platforms, as shown in Figure 5.1-7. You may download as many templates as needed—the download screen remains open until the **Finished** key is pressed.

**Please Select the Appropriate  
Sample Specific Information Template**

**Human Samples:** ☐ Affymetrix Platform  
☐ Spotted Array Platform

**Rodent Samples:** ☐ Affymetrix Platform  
☐ Spotted Array Platform

**Other Samples:** ☐ Affymetrix Platform  
☐ Spotted Array Platform

**Figure 5.1-7 The Download Templates Dialog**

Each template is an Excel spreadsheet with predefined columns specifying the information that should appear in that field. Use Excel to enter the required information and save the file as a tab-delimited text file named *sample.txt*. Figure 5.1-8 shows the Sample-Specific Information Template for the Affymetrix platform and the species *human*.

	A	B	C	D	E	F	G	H	I
1	File Name	Sample Name	Tissue Name	Cell Line Name	Cell Type	Sex	Age	Disease State	Disease Stage
2									
3									
4									
5									
6									
7									
8									
9									
10									

**Figure 5.1-8 Sample Information Template (Affymetrix platform, species human)**

Annotation information must be provided for each hybridization (array). The “File Name” field in the template is the processed data file name *including* the file extension. For Affymetrix files, you should list all cel-files and txt-files (if you are submitting txt-files). Other Affymetrix file types (\*.rpt, \*.dat, \*.chp) need not be included on the Sample-Specific Information

Template. For GenePix experiments, \*.gpr files should be listed, and for UCSF SPOT, the \*Sproc.txt and \*Spot.txt files should be included.

You should always fill in as much sample information as possible. If the requested information is either not available or not applicable to the experiment, those columns can be left blank. After entering information for all of the arrays, remember to save the spreadsheet as a tab-delimited text file named *sample.txt*.

**Step 2.** The second step of the submission process begins by logging into your GEDP user account and selecting the *Submit Experiments* tab. The first page in the submission process (Figure 5.1-9) collects general experiment information. All fields on this page are required.

Enter General Experiment Information

\*Type of Experiment:  ?

\*Experiment Title:  ?

\*Experiment Description:  ?

\*Species:  ?

\*Sample Type:  ?

\*Platform Type:  ?

\*Visibility of Experiment:  ?

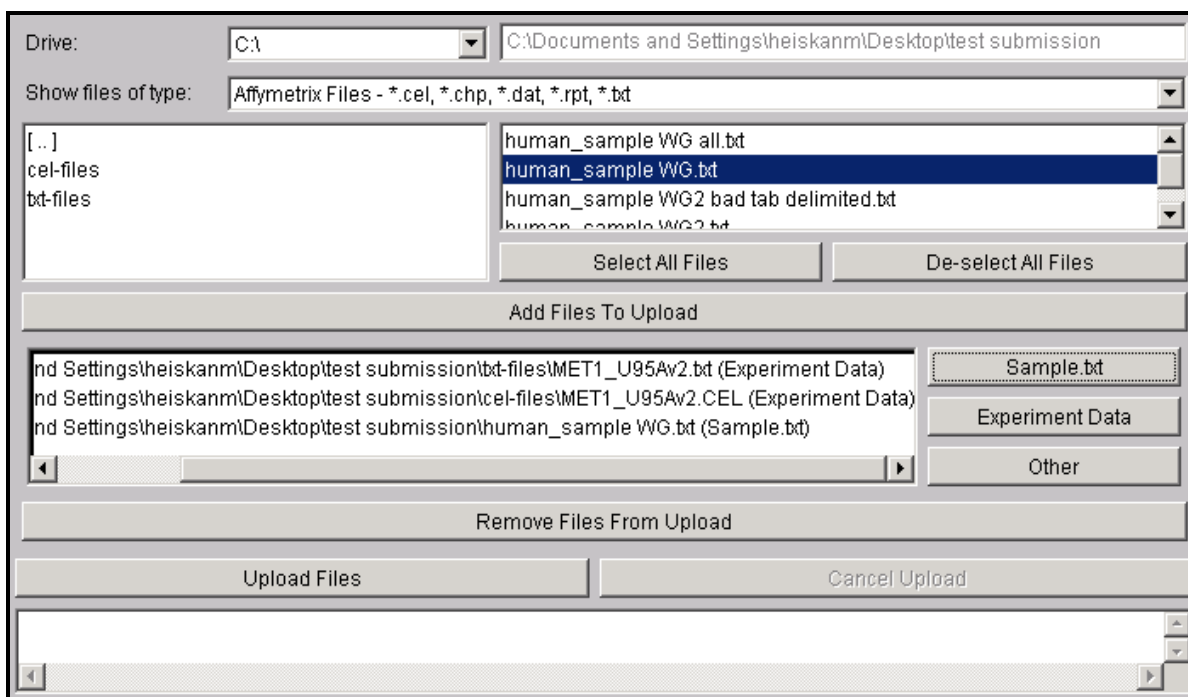
[previous](#) [continue](#) [reset](#)

\*denotes a required field

**Figure 5.1-9 General Experiment Information Form**

**Step 3.** After completing the form, press the **Continue** key to proceed to the Data File Upload page where the applet shown in Figure 5.1-10 will guide you through the upload process. Three types of data files are accepted for upload:

- *Sample*: This is the *sample.txt* file you created in Step 1 using the Sample-Specific Information Template. This file must be included with all submissions.
- *Experiment Data*: These files include the platform-specific data files, i.e., .cel, .txt, .rpt, .dat, and .chp files for Affymetrix; .gpr and .gal files for GenePix spot arrays.
- *Other*: This category is a catch-all for any files not defined by the first two categories. Use “Other” to submit unsupported data files such as ScanAlyze, as well as for any files providing supplemental information—such as analysis results or clinical information. These files will be available for download only.



**Figure 5.1-10 The File Upload Applet.**

The file applet is sensitive to the information that was provided on the previous page. Thus, if an Affymetrix platform was specified in the General Experiment Information, the applet will look for Affymetrix file extensions (.cel, .txt, .rpt, .dat, .chp). Alternatively, if Gene Pix was selected, the applet will look for GenePix file extensions (.gpr, .gal). This information is shown in the box labeled **Show Files of Type**.

Typing in the drive and path to the directory where your experimental files are located will cause all files with the appropriate extensions to be listed in the scrollable selection box above the buttons labeled **Select All Files** and **De-select All Files**. Individual files can be designated for upload by selecting those files and clicking on **Add Files to Upload**.

The files currently selected for upload are listed in the box just beneath the **Add Files to Upload** button, with the file type displayed in parentheses after each file. Files can be removed from this list by selecting the file and pressing the **Remove Files From Upload** button. The file type buttons on the right side can be used to change the specified type for selected files. It is particularly important that the *sample.txt* file is correctly categorized as file type "Sample.txt."

After verifying that the selected files are correct and complete, click on the **Upload Files** button to start the upload. Progress can be monitored on the status bar. When the upload is completed, a message will be displayed stating that all files uploaded successfully.

**Step 4.** After completing the upload, press the **Continue** key to proceed to the next page. This page (Figure 5.1-11 ) allows users to revise the previously submitted experiment information and to add supplemental information specified in the MIAME standard. Filling in these forms is strongly recommended but not currently required. Important information can be included here concerning hardware, software, protocols, and publications. Once you have completed these forms, click on the **Submit Experiment** button to complete the submission.

**NATIONAL CANCER INSTITUTE The Gene Expression Data Portal**  
 GEDP.nci.nih.gov

Home Search Experiments Submit Experiments Analysis Tool Suite Logout

### Add or Modify Experiment Information and Submit

<b>Revise Experiment Information</b>
<a href="#">revise</a> Click to Revise General Experiment Information
<b>Revise Common Sample Information</b>
<a href="#">revise</a> Click to Revise Information Common to All Samples (Arrays)
<b>Hardware</b>
<a href="#">add</a> Enter Hardware Information Pertaining to Scanner or Fluidics
<b>Publications</b>
<a href="#">add</a> Enter publications that are associated with the model. Please provide the PMID (PubMed Identifier) number if a PubMed record exists. This will allow us to link to the abstract of the publication.
<b>Software</b>
<a href="#">add</a> Enter Information about the Software Utilized in Experiment. There are Four Categories: Scanning, Image Processing, Data Mining, and LIMS
<b>Protocols</b>
<a href="#">add</a> Enter Information Regarding a Protocol Used in the Experiment

[Submit Experiment](#)

**Figure 5.1-11 Final Data Submission Page**

A page similar to that shown in Figure 5.1-12 will appear on your screen upon completion of the submission process. Be patient, as it typically takes at least 30 minutes for all of the data to be saved and for the MAGEML document to be generated. The experiment will become available online immediately upon completion.

**NATIONAL CANCER INSTITUTE The Gene Expression Data Portal**  
 GEDP.nci.nih.gov

Home Search Experiments Submit Experiments Analysis Tool Suite Logout

### Your Experiment Submission is Complete....

**Thank You**

Experiment id: 492

Depending upon the number of data files uploaded, and the platform type, saving the experimental data to the database and generating the MAGEML document normally requires approximately 30 minutes. Accordingly, please allow for **at least 30 minutes** prior to searching for a recently submitted experiment.

**Figure 5.1-12 Experiment Submission is Complete Page**



### 5.1.3 XpressionWay

XpressionWay allows you to visualize microarray data from Affymetrix<sup>21</sup> and GenePix platforms—as well as from SAGE (Serial Analysis Of Gene Expression) libraries—in the context of molecular pathways. These pathways are provided courtesy of BioCarta ([www.biocarta.com](http://www.biocarta.com)) and the information is updated approximately every three months.

Figure 5.1-13 outlines the steps required to generate pathway diagrams for experiments you wish to visualize.

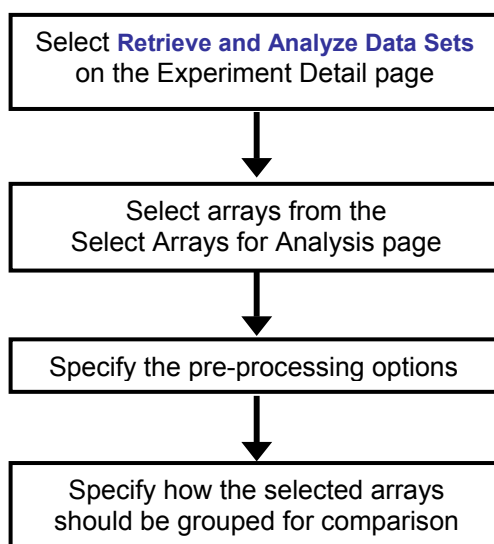


Figure 5.1-13 Flow chart for generating pathway diagrams

**Step 1.** As noted in Section 5.1.1, the Experiment Detail screen (Figure 5.1-5) provides a radio button to **Retrieve and Analyze Data Sets with Array Design**. Selecting this radio button and pressing the **Submit** key will generate a page for selecting data files associated with that experiment. Figure 5.1-14 shows the Select Arrays for Analysis page.<sup>22</sup>

**Step 2.** Initially all files are selected; clicking the **reset** button will unselect all files and allow you to select only those you would like to visualize. Because XpressionWay performs *comparative* analysis, a minimum of two files must be selected. Choose two or more arrays for analysis by selecting the appropriate check boxes and press **next** to set the preprocessing options.

**Step 3.** Different preprocessing options are available for the two platforms. Figure 5.1-15 shows the options available for Affymetrix platforms. The **Spot Filtering** box allows users to specify a threshold for filtering raw intensities. Additional options in this box indicate how that threshold should be interpreted and which probe sets should be excluded—as per the choice made in the **Filter by Absolute Cell** option.

<sup>21</sup> Additional file preparation may be required for Affymetrix data, as described at the end of this section.

<sup>22</sup> Selecting experiments performed on platforms other than Affymetrix or GenePix will generate an error page, as other file types are not currently supported in XpressionWay.

**The Gene Expression Data Portal**  
[Home](#) [Search Experiments](#) [Submit Experiments](#) [Analysis Tool Suite](#) [Login](#)

Select Arrays for Analysis Data Set

Select	File Name	Sample Name	Tissue	Cell Line	Disease Stage	Disease State	Grade	Histology
<input checked="" type="checkbox"/>	BetaGal_Exp2			NIH 3T3				
<input checked="" type="checkbox"/>	BetaGal_Exp1			NIH 3T3				
<input checked="" type="checkbox"/>	E2F1_Exp1			NIH 3T3				
<input checked="" type="checkbox"/>	E2F1_Exp2			NIH 3T3				

[previous](#) [next](#) [reset](#)

Figure 5.1-14 Select Arrays for Analysis Page

The **Scaling/Normalization** box is used to specify how the points should be normalized. The first slot specifies that either the median or the mean intensity—computed over all probe sets—should be used in the normalization. Ratios are then computed and rescaled so as to bring either the mean or median value to 1.0. The mean (or median) of each array is then scaled to the target intensity specified in the third slot. A small percentage of extreme values (as specified by **trim**) may also be excluded when calculating the mean or median.

**Data Set Preprocessing Options**

**Spot Filtering**

Intensity Threshold:

Consider Intensity under Threshold as:

Filter by Absolute Call:

Consider Excluded Probe Sets as:

**Scaling/Normalization**

Scale/Normalize by:

Trim:

Target Intensity:

[previous](#) [next](#)

Figure 5.1-15 Preprocessing Options for Affymetrix Data



Finally, the values observed to be under the specified threshold are adjusted to the threshold or categorized as missing, depending on what was selected for the second option in the **Spot Filtering** box.

Figure 5.1-16 shows similar options for GenePix cDNA platforms. The raw intensity data from cDNA arrays can be filtered by an intensity threshold and/or flags. The ratios are then computed and rescaled for the mean (or median) for all of the selected genes to be 1.0.

**Data Set Preprocessing Options**

**Spot Filtering**

Filter by Flag ☒

Intensity Threshold for Channel(s) either >= 100

Consider Intensity Under Threshold as Threshold

**Normalization**

Normalize by Mean ratio of all genes

previous
next

**Figure 5.1-16 Preprocessing Options for GenePix Data**

**Step 4.** The next step involves choosing the array(s) that will be included in each group. Figure 5.1-17 shows an example screen for defining these groups. If only two arrays were chosen in step 2, then you will only be able to do a one-to-one comparison—select one array for each group. For a group-to-group comparison, select *at least two* arrays for each group. Select **next** to view the list of pathways represented in the array data you have selected. Clicking on a listed pathway will generate the pathway viewing screen for that path.

**Select Samples/Arrays for Comparison**

One-to-one Comparison (Select ONE for each group)

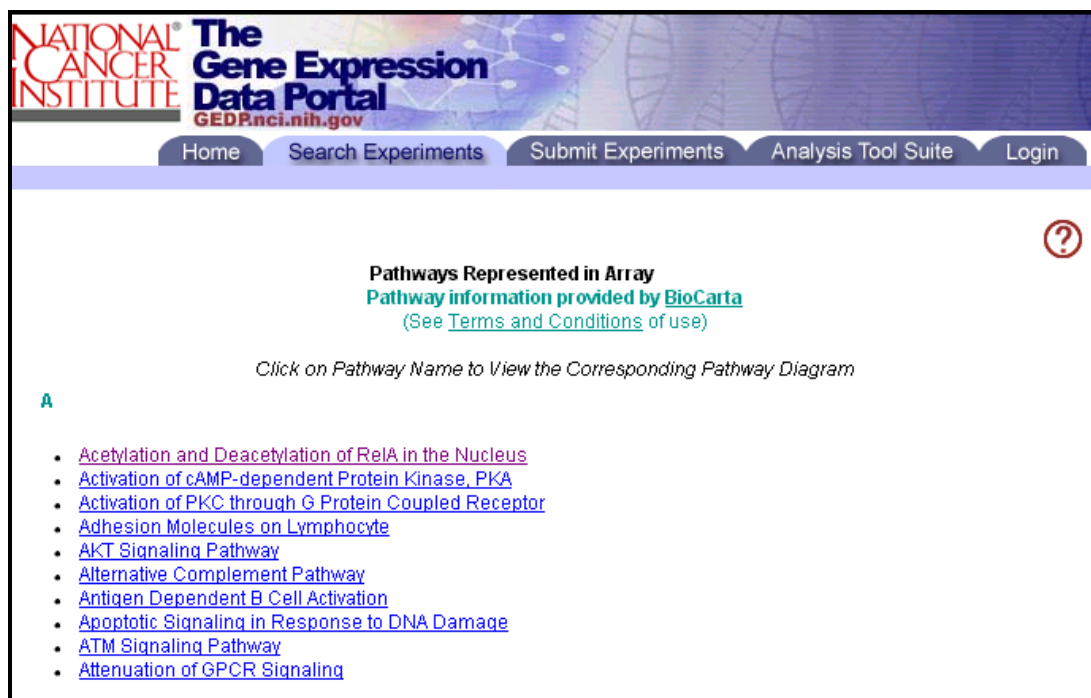
Group-to-group Comparison (Select at least TWO for each group)

Array Name	Group A	Group B
E2F1_Exp1	<input checked="" type="radio"/>	<input type="radio"/>
BetaGal_Exp1	<input checked="" type="radio"/>	<input type="radio"/>
BetaGal_Exp2	<input type="radio"/>	<input checked="" type="radio"/>
E2F1_Exp2	<input type="radio"/>	<input checked="" type="radio"/>

reset
previous
next

**Figure 5.1-17 Defining How the Arrays Will be Grouped for Comparison**

Figure 5.1-18 shows a sample list of pathways for viewing. Clicking on a pathway name diagram will display a Scalable Vector Graphics diagram for that pathway.<sup>23</sup>



**Figure 5.1-18 Display Page for Pathways Represented in the Array**

For example, the pathway diagram in Figure 5.1-19 was generated by clicking on the AKT Signaling Pathway. Detailed information about various genes in the pathway can be found on the Gene Information pages, which are hyperlinked to the gene names in the diagram.

Buttons to the right of the diagram can be used as follows:

- Pressing **Genes On Chip** highlights all of the genes included on the chip in pink. In general, there is not a one-to-one correspondence between genes occurring on the displayed pathway and those arrayed on the chip: some genes on the chip may not be included on any given pathway, and all of the genes on a particular pathway may not be included in the chip
- The **Expression +/-** button can be used to highlight those genes showing a two-fold or greater change in expression when the two groups (defined in Step 4) are compared with one another. Up-regulated genes are highlighted in red and down-regulated genes are highlighted in blue. Genes shown in green represent genes with multiple features on the chip and inconsistent results.
- Pressing **Reset** resets the pathway diagram to its initial state.
- **Pathway Summary Report** generates a summary of the comparative expression analysis used for the display of up- and down-regulated genes in the diagram.

<sup>23</sup> Scalable Vector Graphics (SVG) is a language for describing two-dimensional vector and mixed vector/raster graphics in XML. First-time users will be prompted to download the Adobe SVG Viewer 3.0, which is available on Adobe's Web site free of charge (<http://www.adobe.com/svg/viewer/install/main.html>).

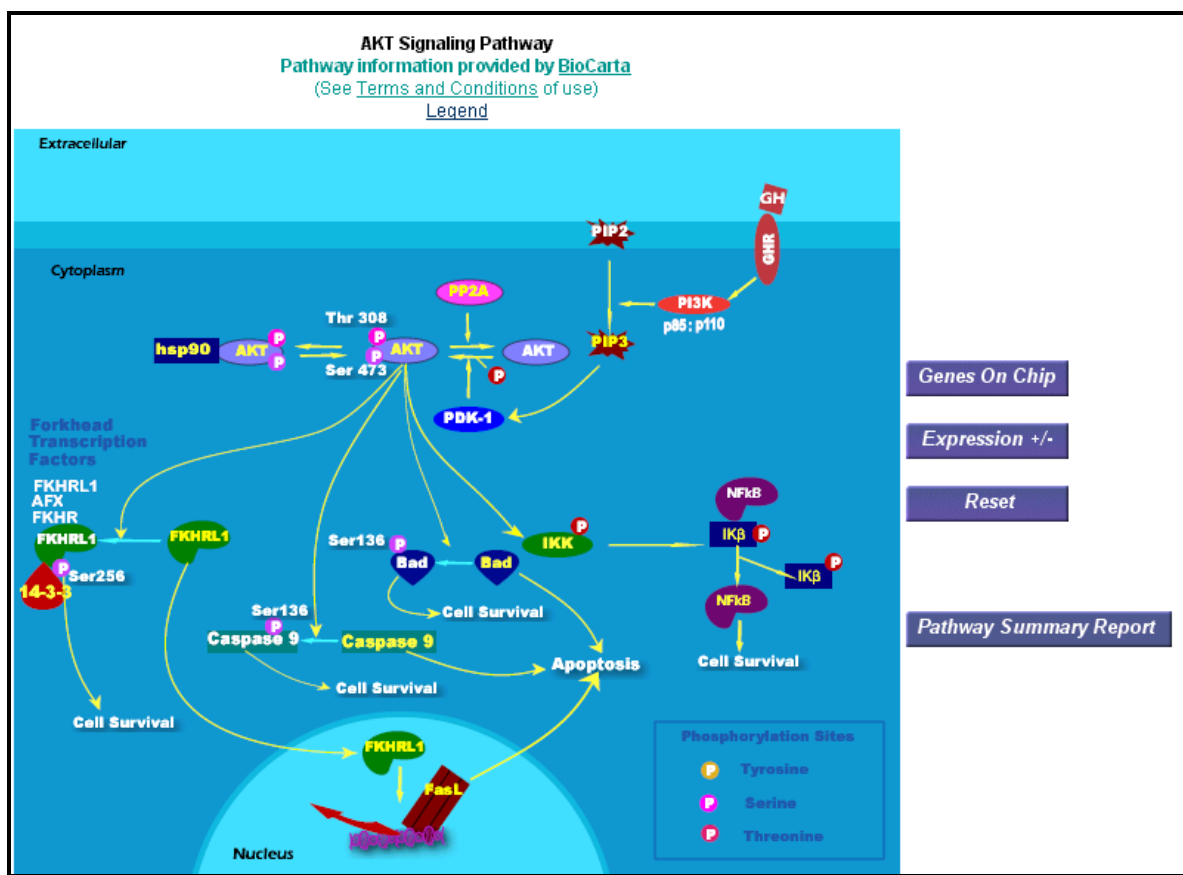




















Figure 5.1-19 SVG Diagram for the AKT Signaling Pathway

Figure 5.1-20 shows one such sample summary report generated for Affymetrix data. The summary report lists all the reporters (probes), their mean intensity values, and the target genes. As shown in Figure 5.1-20, each target gene occupies one row in the summary table.

Pathway Summary Report						
Target Gene	Reporter Name	Mean of Group A		Mean of Group B		P<
<a href="#">SHC1</a>	AA203501_at	311.7			199.5	0.24551
<a href="#">PTPN11</a>	C02549_at	137.1			78.2	0.14220
<a href="#">MAPK8</a>	L40392_at	797.1			712.8	0.68150
<a href="#">PTPN11</a>	RC_AA052953_at	81.0			66.3	0.62575
<a href="#">HRAS</a>	RC_AA059006_at	24.0			26.3	0.67448
<a href="#">MAPK8</a>	RC_AA180321_at	61.8			79.0	0.56195
<a href="#">PTPN11</a>	RC_AA259147_at	926.1			1016.1	0.18836
<a href="#">PIK3CG</a>	RC_AA448663_at	77.9			80.6	0.92594
<a href="#">MAPK8</a>	RC_AA463692_s_at	230.0			231.2	0.99229









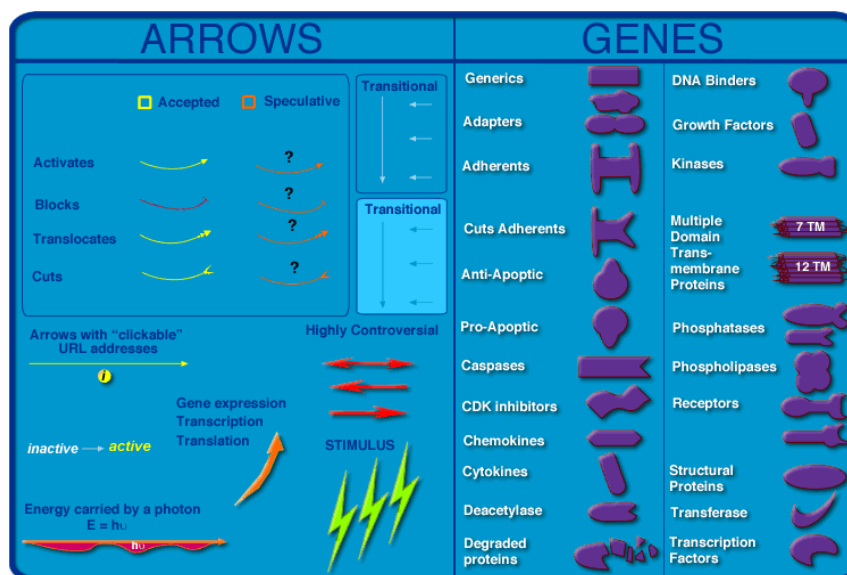
Color Code									
log <sub>2</sub> (intensity) represented in number of S.D. from the mean	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2

Figure 5.1-20 Pathway Summary Report for Affymetrix Data

The color-coding is derived as a function of the number of standard deviations the observed numerical values are from the overall mean observed for that gene, across all selected arrays using a log<sub>2</sub>-scale. The last column of the summary table displays a two-sample t-test p-value, to indicate the significance of the observed difference between the two groups. The color-coding scheme is displayed separately just beneath the summary table.

A similar table layout is used for cDNA microarrays in summarizing the expression analysis. Again, a log<sub>2</sub> color-coding scheme is used to graphically represent the observed values. In this case, however, intensity ratios—rather than direct signals—are reported numerically and encoded graphically by the log<sub>2</sub>-based color codes. Like the gene symbols in the pathway diagrams, each gene name in the summary tables is hyperlinked to the *Gene Info* page for that gene.

Figure 5.1-21 shows the Legend screen for pathway diagrams, which can be invoked by pressing on the Legend key near the top of the diagram display.



**Figure 5.1-21 Pathway diagram legend**

It is also possible to apply the XpressionWay visualization tool to CGAP SAGE data as well as to private data uploaded by the user for this purpose. Selecting the **Analysis Tool Suite** tab on the GEDP home page provides access to these additional options.

Additional information about the XpressionWay tool is available from the online help that can be accessed by pressing on the [? icon](#). The next page summarizes the steps required to prepare an Affymetrix data set for XpressionWay analysis when the .txt and .rpt files are not available.

## Notes on Generating Affymetrix Data Files for Pathway Visualization

Analysis of Affymetrix data requires that you generate \*.txt and \*.rpt files using the Affymetrix MAS5 software as outlined below.

### Expression Analysis Settings:

Scaling: All probe sets (Target signal 500)

Normalization: user defined, 1

No probe mask

No base line

Parameters: Affymetrix default-settings

### Generating the \*.txt files:

1. Perform an "absolute expression analysis" on the \*.cel file, as opposed to "comparison expression analysis (use a baseline)."
2. To generate the text file in the format that we are currently using:
  - Select "Analysis" from the pull-down menu and click on "Options";
  - Click on the "Metrics" tab in the option panel;
  - Check "Save All Metrics Results";
  - Select "Save Each Analysis to a Separate File";
  - Click "OK" to close the pop-up window.
3. In the data window click on the "Metrics" tab to view the data in metrics format. Click "Save" to generate the \*.txt data file.

### Generating the \*.rpt files:

Right-click on a chip file and select report.

The file is saved automatically to the mapped folder.

**Figure 5.1-22 Generating Affymetrix data files**

#### 5.1.4 The Affy Cel File Analysis Center

Pressing the *Analysis Tool Suite* tab on the GEDP home page also provides access to the Affy Cel File Analysis Center, as shown in Figure 5.1-22.

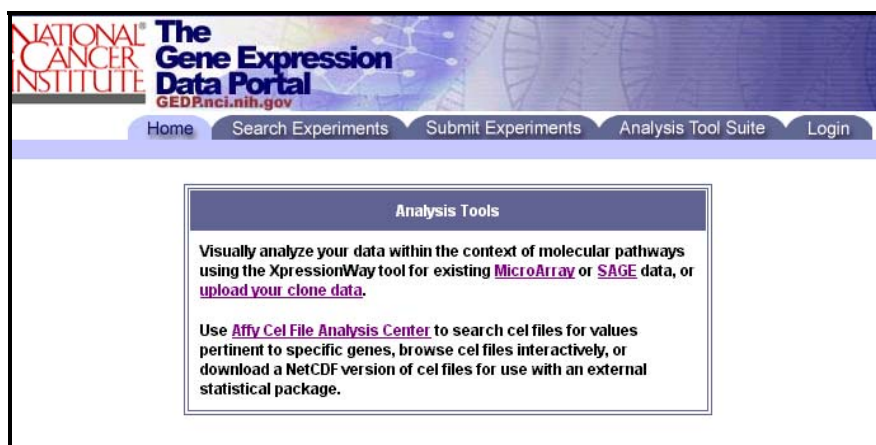


Figure 5.1-23 GEDP Analysis Tools

The Affy Cel File Analysis Center can be used to view Affymetrix data at probe level and to download network Common Data Form (NetCDF) versions of cel files. Clicking on the [Affy Cel File Analysis Center](#) hyperlink generates a new page listing the Affymetrix data sets in GEDP that are available for this type of analysis. An example listing is shown in Figure 5.1-23.

Available Experiments				
Exp ID	Investigators Name	Experiment Title	Platform Type	Array Design (s)
<a href="#">231</a>	Hanash, Samir	Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis	Affymetrix	Hu6800
<a href="#">273</a>	Livingston, David	Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses	Affymetrix	HG_U95Av2
<a href="#">308</a>	Hanash, Samir	Ovarian Tumors from U. Michigan on Affymetrix HuFL chips	Affymetrix	Hu6800
<a href="#">311</a>	Hanash, Samir	Lung adenocarcinomas and normals from U. of Michigan on Affymetrix HuGeneFL arrays.	Affymetrix	Hu6800

Figure 5.1-24 Experimental Data Available for Affy Cel File Analysis

Clicking on an experiment ID in the above table brings up a search page listing the cel files associated with that experiment. Figure 5.1-24 shows an example Cel File Search page. Clicking on a filename in the listing initializes the download of that cel file in binary netCDF format; selecting the associated [Browse Data](#) hyperlink opens that file for browsing.

A text-based searchbox at the top of the search page is provided for entering UniGene symbols for the genes of interest. When searching for multiple genes, these UniGene symbols should be separated by commas or by spaces. Typing in these symbols and pressing the [Search Cel Files](#) button will initialize a search for those genes in all of the listed cel files.

**The Gene Expression Data Portal**  
[Home](#) [Search Experiments](#) [Submit Experiments](#) [Analysis Tool Suite](#) [Login](#)

**Affy Cel Lookup Experiment: 955**

Type in a space or comma delimited list of gene symbols and then press the button below. The search may take a few seconds. When it is complete the values from the experiment cel files that are pertinent to the specified gene(s) will be displayed.

Gene Symbols:

Available Cel Files	
<a href="#">chtn-oe-069.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-oe-080.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-oe-048.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-009.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-116.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-081.cel</a>	<a href="#">Browse Data</a>
<a href="#">cu-os-004.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-008.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-072.cel</a>	<a href="#">Browse Data</a>
<a href="#">chtn-os-098.cel</a>	<a href="#">Browse Data</a>
<a href="#">um-os-010.cel</a>	<a href="#">Browse Data</a>

**Figure 5.1-25 The Affy Cel File Search Page**

The Search Results table (Figure 5.1-25) summarizes the data for genes matching the specified UniGene symbols. The two leftmost columns report the probe cell x and y coordinates in cell units. The next three columns show the mean probe cell intensity, the standard deviation from the mean observed for that probe, and the number of pixels included in the calculation.

Probe cells with non-uniform intensity are automatically identified by Affymetrix MAS5 software and are marked as outliers. All masked outliers will be excluded from further analysis.

Affy Cel Search Results for Experiment: 955						
myc - chtn-oe-069.cel						
X	Y	Mean	Std Dev	Num Pixels	Is Outlier	Is Mask
168	441	1350.5	92.2	30	false	false
168	442	315.3	44.3	36	false	false
169	442	520.0	49.1	36	false	false
169	441	488.0	55.4	30	false	false
170	441	422.0	53.5	36	false	false
170	442	420.8	84.7	30	false	false
171	441	697.8	62.3	36	false	false
171	442	456.5	71.5	36	false	false
172	441	149.3	30.1	36	false	false
172	442	124.8	19.8	36	false	false
173	441	182.3	129.9	36	false	false

**Figure 5.1-26 Partial Listing of an Affy Cel Search Results Page**



## 5.2 caWorkbench

caWorkbench provides a comprehensive and extendible suite of desktop software tools that can be applied to the analysis, visualization, and annotation of microarray data. In addition to the analysis and visualization tools routinely found in microarray software tools today, caWorkbench provides an enhanced environment via its integration with the Cancer Bioinformatics Infrastructure Objects (caBIO)<sup>24</sup>. This integration provides caWorkbench users with access to publicly available microarray data on a remote NCI server; to the CGAP web site's gene annotation pages, and to the pathway visualization diagrams generated by BioCarta. This last capability allows users to view the observed microarray data in the context of metabolic and signal transduction pathways.

This first release of caWorkbench—also referred to as the Gene Expression Analysis Workbench (GEAW)—represents a growing collection of freely available desktop tools built on open source technology. The workbench is intended to support a variety of the input formats in which microarray data are found; its open-ended design supports the extension of the software to accept additional formats as needed.

The present version of caWorkbench supports both Affymetrix (.txt, MAS 4.0/5.0) and GenePix (.gpr) files. A simple plug-in framework allows users to further define and use any input format they wish. Similarly, this plug-in framework supports the addition of any number of user-defined filters, normalizers, and analysis algorithms.

The [NCICB download center](#) provides links for downloading caWorkbench as well as for registering users to receive news and ongoing notifications about upcoming revisions and future releases. The current release includes a simple on-line help facility that will be extended as the software matures.

This chapter provides an overview of a rather complex software suite, and assumes that the user has some experience with microarray data analysis. A more in-depth tutorial is beyond the scope of this manual. The discussion which follows outlines procedures for downloading and installing the software; for loading data files; for using visualizations; and for annotating data.

### 5.2.1 Downloading, installing and starting caWorkbench

caWorkbench is a stand-alone application; the only system requirement for installation is that your computer must contain a Java runtime environment (JRE) or development kit (SDK) of version 1.4.0 or higher.

To download caWorkbench, begin by pointing your browser to the NCICB download center at <http://ncicb.nci.nih.gov/download/index.jsp>. This page requires that you register your email address, name, and institutional affiliation before actually entering the center. From the center, clicking on the caWorkbench hypertext will redirect your browser to the License Agreement page, where you will also have the opportunity to register with the caWorkbench listserv. It is highly recommended that users register with this listserv. In addition to providing up-to-date information on revisions and new releases, the listserv provides a forum for users to ask questions, report problems, and exchange information.

---

<sup>24</sup> The NCICB Technical Guide provides a detailed description of the caBIO project and its application programming interface (API). Chapter 4 of this manual provides a tutorial description of the web-based interface to caBIO, BIOgopher.



Pressing the **Download** button on the License Agreement page will redirect your browser to the page where you can download the software as a *zip* file. A separate README file is also available from that site, but as this file is included in the zip file, there is no need to download it separately.

The next step is to extract all files from the zip file to a convenient directory such as C:\caWorkbench. As described in the README file, windows users must first edit the file named *geaw.bat* before running the program. Similarly, Unix users should edit *geaw.sh*. Line 12 of this file identifies the location of the Java home directory, and should be modified to reflect that location on your machine.

Windows 98 users will also need to make the following changes to accommodate the Windows 98 DOS shell:

1. Comment out *setlocal* and *endlocal* by inserting “REM” at the beginning of those lines.
2. Remove “-cp %classpath%” from the line near the bottom of the file where it reads  
"%JAVA\_HOME%\bin\java" -Xmx256m -cp %classpath% core.config.UILauncher

Once the .bat (or .sh) file has been modified, you may wish to create a shortcut to that file on your desktop. Clicking on the shortcut icon will then start up caWorkbench.

The README file also contains notes on compiling the software and modifying the application’s properties file. All of the Java source code is included in the download, and a *build.xml* file is included for advanced users who wish to extend and rebuild the application using ANT. No further discussion of this advanced feature is included here.

The properties file is a simple text file called *application.properties*. Two properties you may need to edit are *browser.path* and *java.security.policy*. The former specifies the path to the executable for your browser. Windows users who use a standard installation of Internet Explorer should not need to modify the default value provided. The *java.security.policy* property specifies internet security options and is set, by default, to *java.policy*—a text file that is also included in the download.

## 5.2.2 Layout of the caWorkbench Interface

Figure 5.2-1 shows a screenshot of caWorkbench’s graphical interface. As seen there, the workspace is divided into 4 resizable panels whose functionality is further defined by the folder tabs running across the top of each panel. Each panel can be arbitrarily resized by clicking on an edge of that panel’s frame and dragging the mouse. In addition, the triangular shaped wedges (on the left sides of the horizontal separators and at the top of the vertical separator) can be clicked on to maximize that frame vertically and/or horizontally.

Each of these configurable panels is described in more detail in the sections that follow; the purpose of this section is to provide an overall orientation. Moving from left to right and from top to bottom, these panels include a *Project* window (1), a *View* window (2), a *Marker/Phenotype* window (3), and an *Analysis/Annotation* window (4).

caWorkbench organizes data files using a *workspace/project* paradigm. A project is analogous to a “virtual” folder, as it allows individual data sets to be grouped together without modifying their physical storage locations. Once a project’s data sets have been defined, it is possible to open, save, or close all data sets in that project with a single action.

A typical use of the project facility is to associate multiple data sets from the same experiment with one another in a single project file. In addition to loaded data files, other types of data generated during the session—e.g., images, results from applying filters/normalization, etc.—can also be saved and associated with a particular project.

Multiple projects can, in turn, be managed within a single workspace. A user can create a project in a workspace, delete an existing project from a workspace, or rename a project. The application supports handling an arbitrary number of projects in a single workspace.

In summary, a workspace may contain multiple projects, which may themselves contain a variety of raw, filtered, normalized, or otherwise annotated microarray data sets. Workspaces are saved as files with a *.wsp* extension. Projects can only be saved and/or accessed as part of a workspace.

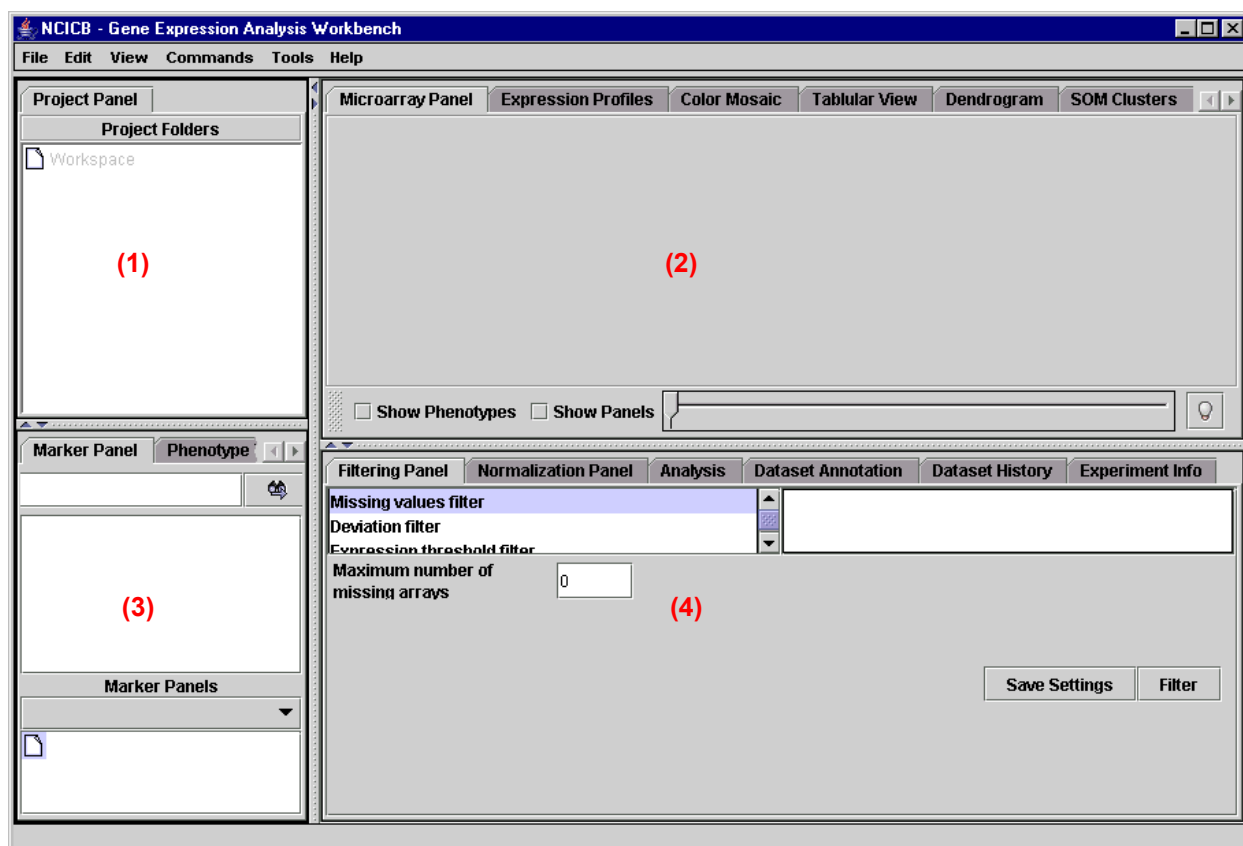


Figure 5.2-1 Layout of the caWorkbench Graphical Interface

In addition to the controls provided within each panel, a main menubar appears at the top of the screen. The first menu option, **file**, is used to manage the opening, creating, deleting, and saving of workspaces, projects, and files. Most of the options included in the main menubar are applicable to the Project and Marker/Phenotype windows, and are described below.

### 5.2.3 The Project Window

The Project window—also referred to as the Project Tree Panel in the online help facility—provides a centralized area for managing projects and files in the current workspace. Operations

on this window are controlled by the **file** and **edit** options in the main menubar. Most of these menu options can also be accessed by right clicking the mouse when it is located over an appropriate element in the Project or Marker/Phenotype window.

Several of the operations in the drop-down **file** menu are not applicable to the Project window. We list these operations here for completeness, but defer any further discussion to the relevant sections.

Figure 5.2-3 illustrates the expanded options for the first three operations under **File**, i.e., **Open**, **Save**, and **New**. These operations can be selectively applied to open, save, or create workspaces, projects, files, and *panel sets*. Panel sets are displayed in the Marker/Phenotype window immediately below the Project window, and are described in [Section 5.2-4](#).

Not all operations are applicable to all four types of elements however. As noted above, projects can only be opened, saved, or created as part of a workspace. Similarly, files can only be opened and/or saved as part of a project, and cannot be explicitly created. Selecting the **File→Open→File** operation without having first defined a project in which to open that file will generate a prompt advising the user to first select a project in the Project window.

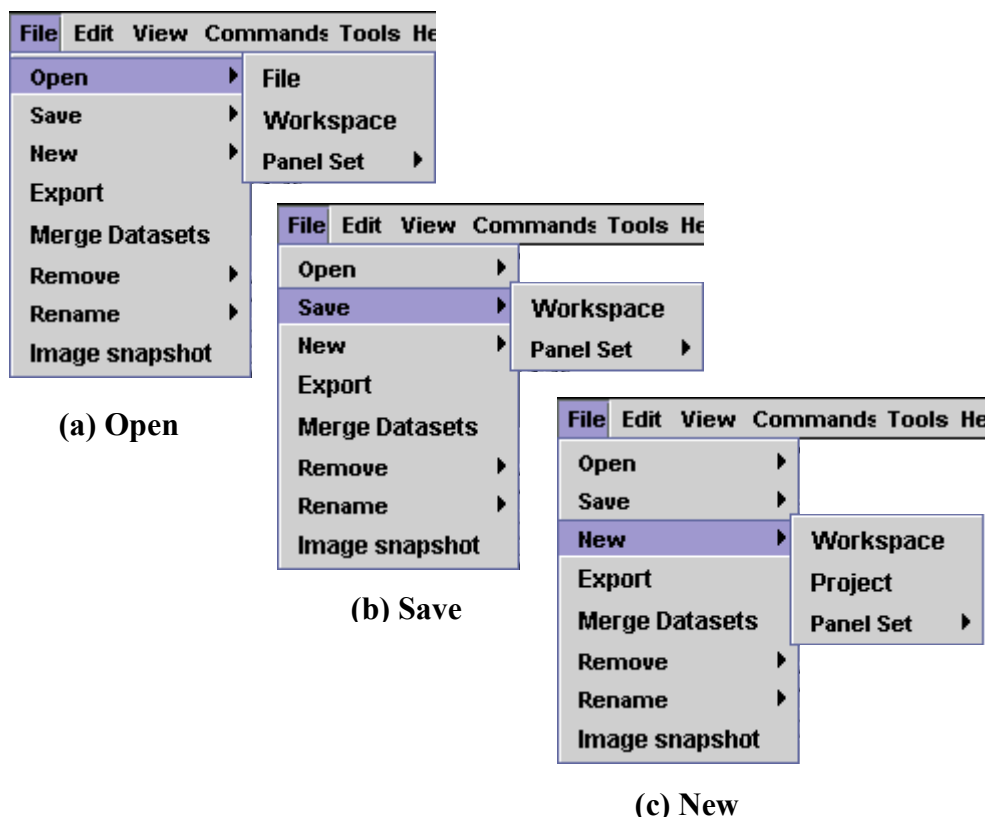


Figure 5.2-2 The Open, Save, and New Operations

When the application starts up, a blank workspace is created by default. A new (originally empty) project can be attached to the workspace by selecting **File→New→Project** from the drop-down menu (Figure 5.2-2(c)). Data files can then be added to the project by selecting that project in the Project window and using the **File→Open→File** option (Figure 5.2-2(a)). The workspace

can be saved in a .wsp file using the **File→Save→Workspace** option and reopened at any future time using the **File→Open→Workspace** option.

The default option in the pop-up "Open File" dialog box is to open a local file, as indicated by the radio button at the bottom of the dialog box in Figure 5.2-3(a). Selecting the appropriate file type from the pull-down list in the local file dialog box will display all files of that type. Double clicking on the selected file from the list of those available will then close the file dialog box and add that file to the current project. A small set of example datasets are available with the download package, in the *example data* directory.

Alternatively, to access remote data sets stored on the NCI server, you must first select the rightmost radio button at the bottom of the dialog box, as shown in Figure 5.2-3(b). As seen there, the lefthand panel lists the files available for selection, and a scrollable text window on the right displays information about the currently selected experiment.

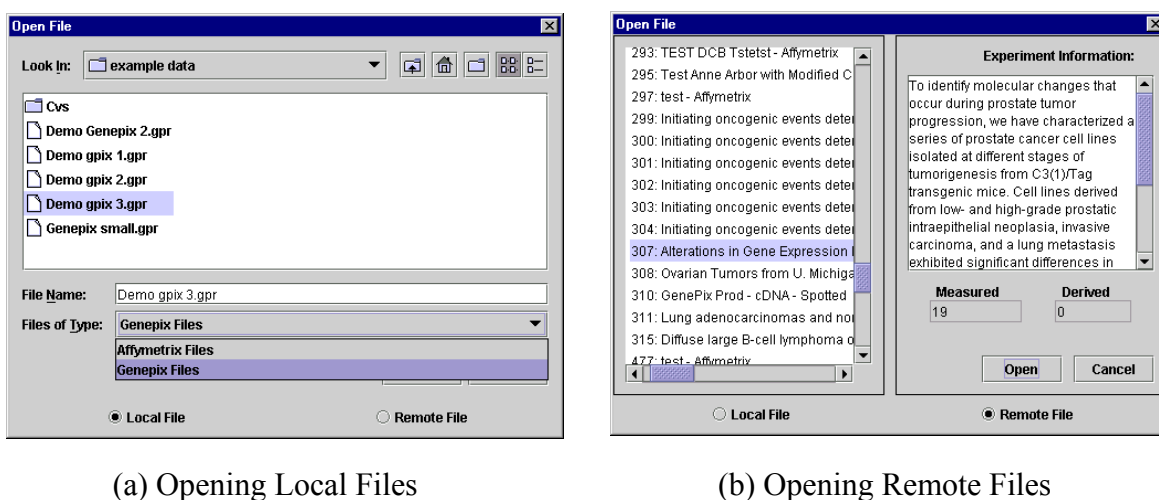


Figure 5.2-3 The Local and Remote Open File Dialog Boxes

To open a remote file, begin by right clicking on the selected experiment in the left panel. A small pop-up text control saying **Get bioassays** will appear; clicking on this will expand the experiment entry to expose a list of the files associated with that experiment, as in Figure 5.2-4. Selecting a file from that list and pressing **Open** in the right panel will import that file to the current project.

In summary, the **File→Open** command can be used to add files to a selected project and to open workspaces and panel sets. Files can be opened from a local disk or from the server at NCI, and are always saved as part of a project. Finally, the operation of opening or creating a new workspace will lose any work that has been done in the current workspace, so be sure to save your work before performing either of these actions.

The **File→Save** option in Figure 5.2-2(b) is applicable to workspaces and panel sets only; files and projects are implicitly saved as constituents of a workspace. The **File→New** option in Figure 5.2-2(c) is only applicable to workspaces, projects and panel sets; "new" data sets can only be derived by performing operations on existing data. These derived data sets will however, be saved as part of the project in which they were derived.

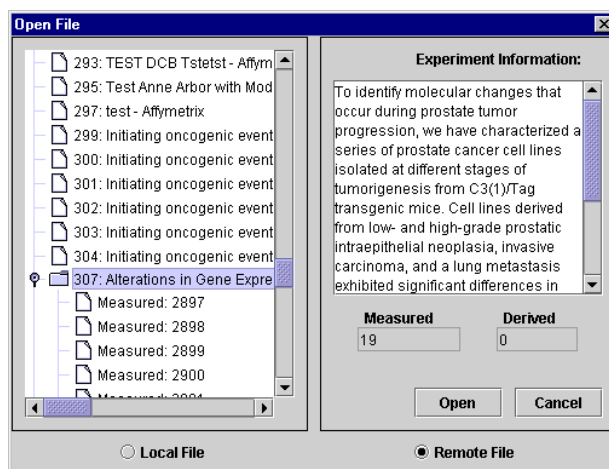


Figure 5.2-4 Opening a Remote File

The next two options in the **File** pull-down menu are **Export** and **Merge Datasets**. The export option can be used to save the data in a data file or a selected image (see [Section 5.2-5](#)) in a new format. Currently, only the “cluster v2.20” export format for text is supported. The **Merge Datasets** operation combines two or more data sets generated using the same array platform to produce a single set. However, *only those data sets included in the same project can be merged*.

Figure 5.2-5 shows the expanded drop-down menus associated with the next two file operations. The **File→Remove** option is used to remove images and files from a project, and projects and marker panels from a workspace. In all cases, the user must first select the object to be removed in the Project window before executing the operation.

The **File→Rename** option is used to rename marker and phenotype panels, as described in [Section 5.2.4](#). It is also possible to rename projects and data files from within the Project window, by right-clicking on that element and using the shortcut pop-up menu. The drop-down **Edit** menu in the main menubar also provides this option.

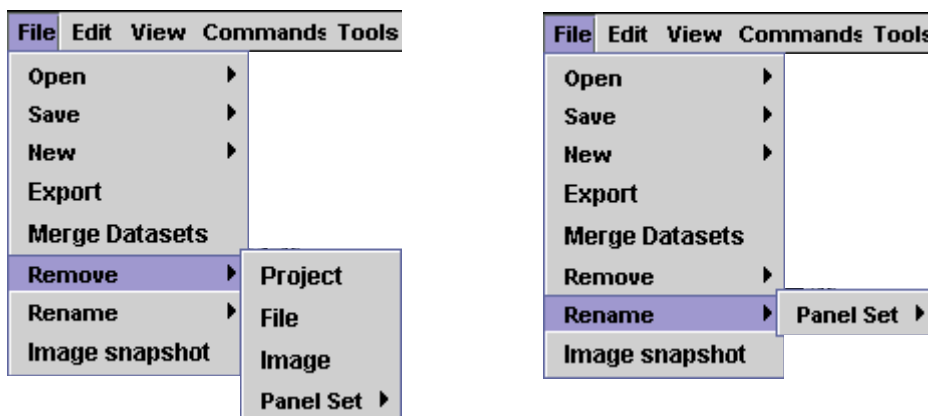


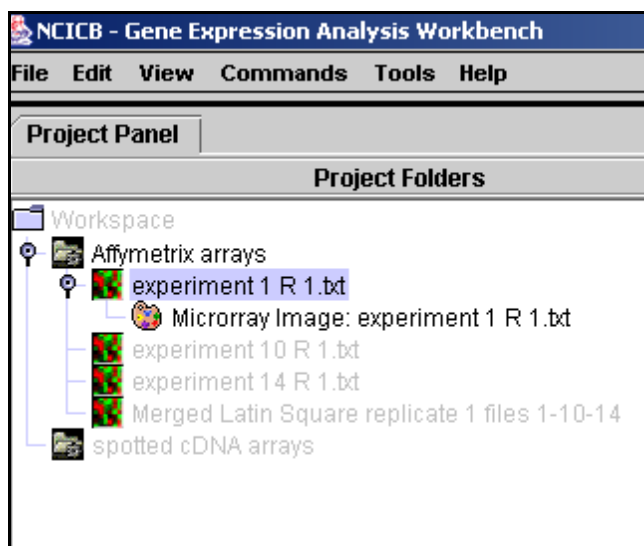
Figure 5.2-5 The Remove and Rename Operations

The last option in the **File** drop-down menu is the **Image snapshot** operation, which is used to take snapshots in the View window and is described in [Section 5.2.5](#). For convenience, all of the file operations are summarized in [Table 5.2-1](#).

**Table 5.2-1 File Operations in the Main Menubar**

Command	Arguments	Action
<b>File→Open</b>	files, workspaces, panel sets	Opens a new workspace, or a file or panel set in the current workspace.
<b>File→Save</b>	workspaces, panel sets	Saves the workspace or panel set.
<b>File→New</b>	workspaces, projects, panel sets	Creates a new workspace, or a new project or panel set in the current workspace.
<b>File→Export</b>	images, files	Saves an image or data set in a new format.
<b>File→Merge Datasets</b>	files	Merges two or more data sets to generate a single combined set.
<b>File→Remove</b>	projects, files, panel sets, images	Removes files and/or images from a project, or projects and/or panel sets from a workspace.
<b>File→Rename</b>	panel sets	Renames objects in the Marker/Phenotype window.
<b>File→Image snapshot</b>	objects in the View window	Takes a snapshot of an object in the View window.

The operations described here all manipulate the workspace, projects, files, and images that are visible in the Project Tree Panel. Only a single workspace can be open at one time, but multiple projects, files, and images can be managed within that workspace. The Project window uses a hierarchical treelike structure to manage these elements, as shown by the example in Figure 5.2-6.



**Figure 5.2-6 An Example Project Tree**

Used to capture a “working session,” the workspace is represented by a folder icon at the very top of the file hierarchy into which all of the data generated during a user session can be

subsumed. Items contained in a workspace include projects, which may themselves contain a variety of raw, filtered, normalized or otherwise annotated microarray data sets.

Multiple projects can be accommodated under one workspace heading. Multiple data sets and derivatives thereof can be grouped within a single project. It is important to note, however, that some operations require data sets to be part of the same project, and in some cases, in the same file. For instance, two microarrays cannot be viewed side by side unless they have been merged into one file. Moreover, two data sets cannot be merged into one file unless they are included in the same project.

## 5.2.4 The Marker/Phenotype Window

### 5.2.4.1 Marker Panel Sets and Individual Marker Panels

The term *marker* has many different interpretations in various contexts. caWorkbench uses the term marker to refer to a gene probe. The definition of what constitutes a gene probe in turn depends on the type of microarray platform. On Affymetrix platforms, gene probes are oligonucleotides synthesized *in situ*. On other platforms (e.g. GenePix), gene probes are oligonucleotides or cloned DNA fragments deposited and immobilized on the substrate by various techniques.

A *marker panel* is a user-defined grouping of several gene probes. Typically, these marker panels probe for genes of specific interest, of importance for certain disease or developmental states, or for characteristic changes in gene expression that may be hallmarks of a tumor in a particular tissue.

A master list of all gene probes in the currently selected data set is displayed under the Marker Panel tab, and a color-coded image of the corresponding gene expression measurements is shown in the View window's Microarray Panel. Individual or groups of gene probes can be selected from this master list and added to smaller marker panels for use in a specific study. The smaller window (*Marker Panels*) immediately below the master list provides an area where selected gene probes can be examined more closely via user-defined marker panels.

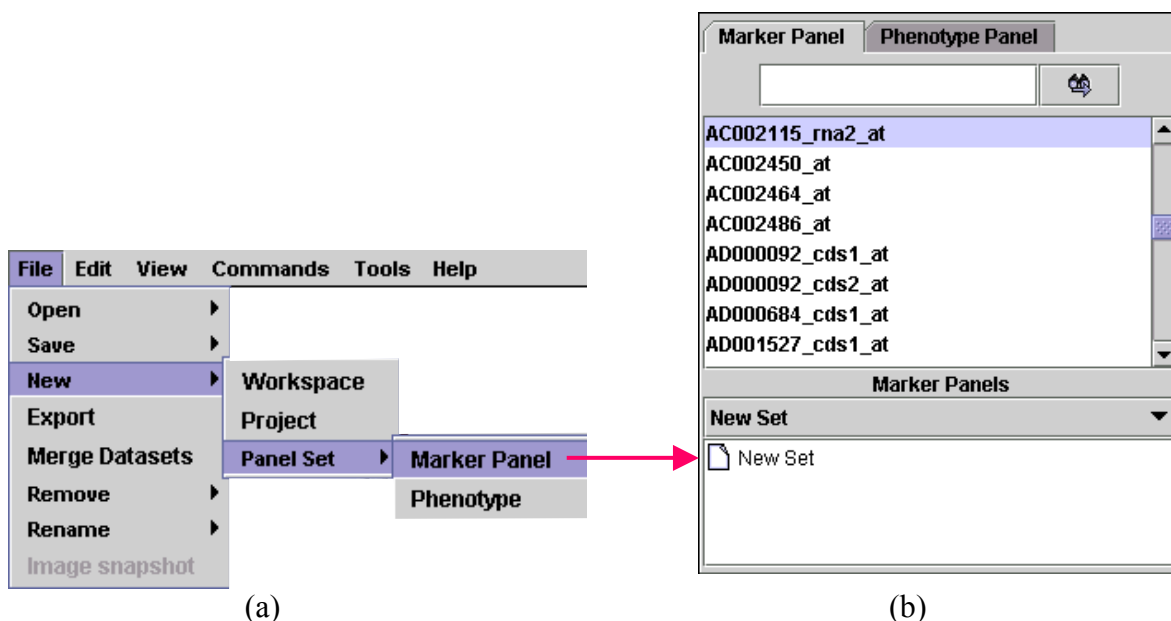


Figure 5.2-7 Creating a New Marker Panel Set



All marker panels must belong to a *set* of marker panels, so the first step in creating a panel is to create a marker panel set. A new marker panel set is created by selecting **File→New→Panel Set→Marker Panel** (Figure 5.2-7(a)). The new marker panel set is then displayed in the Marker Panels pane in the bottom-left corner of the screen as “New Set” (Figure 5.2-7(b)).

Initially, the new panel set is empty. Adding a new panel to the set *and* a marker to the new panel can be done in one step by right clicking on a gene probe in the master list and selecting **Add to Panel** from the pop-up menu. Like the Project window, the Marker Panels window uses a tree-structured hierarchical representation of its panel sets. As a result of adding the first probe to the empty panel set, the Marker Panels window now contains a “New Panel,” as shown in Figure 5.2-8(a). Clicking on the key-like icon next to the panel “opens” it, and shows that the selected probe is now contained in that panel (Figure 5.2-8(b)).

It is also possible to add markers to a panel *without* explicitly creating a new panel set *or* a new panel. Using the **Add to Panel** command from the pop-up menu will automatically create a new default panel set (if none are defined) and a new default panel—if none is selected.

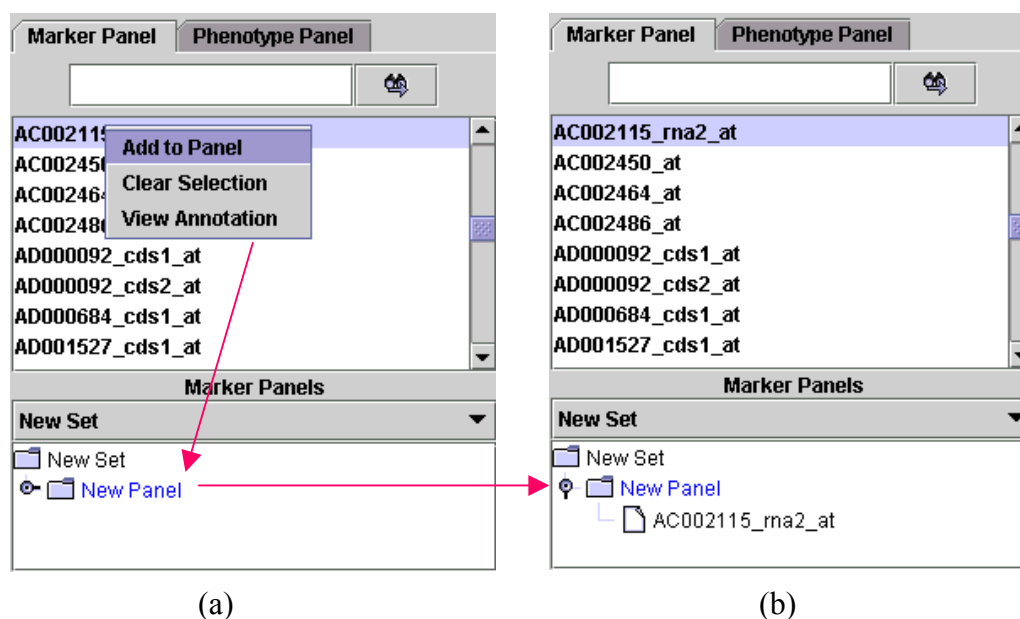


Figure 5.2-8 Adding a Probe to a Marker Panel

Additional probes can be added to the same panel by first selecting that panel in the Marker Panels window, and then selecting the probe from the master list, right clicking the mouse, and selecting **Add to Panel** from the pop-up menu. If this action is executed *without* first selecting a panel, a new panel will be created to hold the selected probe and added to the marker set

In addition to having multiple panels within one panel set, it is also possible to have multiple panel sets within a workspace. The steps depicted in Figure 5.2.8 can be repeated any number of times to produce multiple panel sets.

As soon as a specific microarray is selected in a project folder, the entire complement of markers on that array is displayed in the Marker/Phenotype Window under the *Marker Panel* tab. For example, in Figure 5.2-9, the user has just created a new project and loaded a single file—*aml8\_965\_ab\_hu68-300markers.txt* (from the *example data* directory)—to that project.

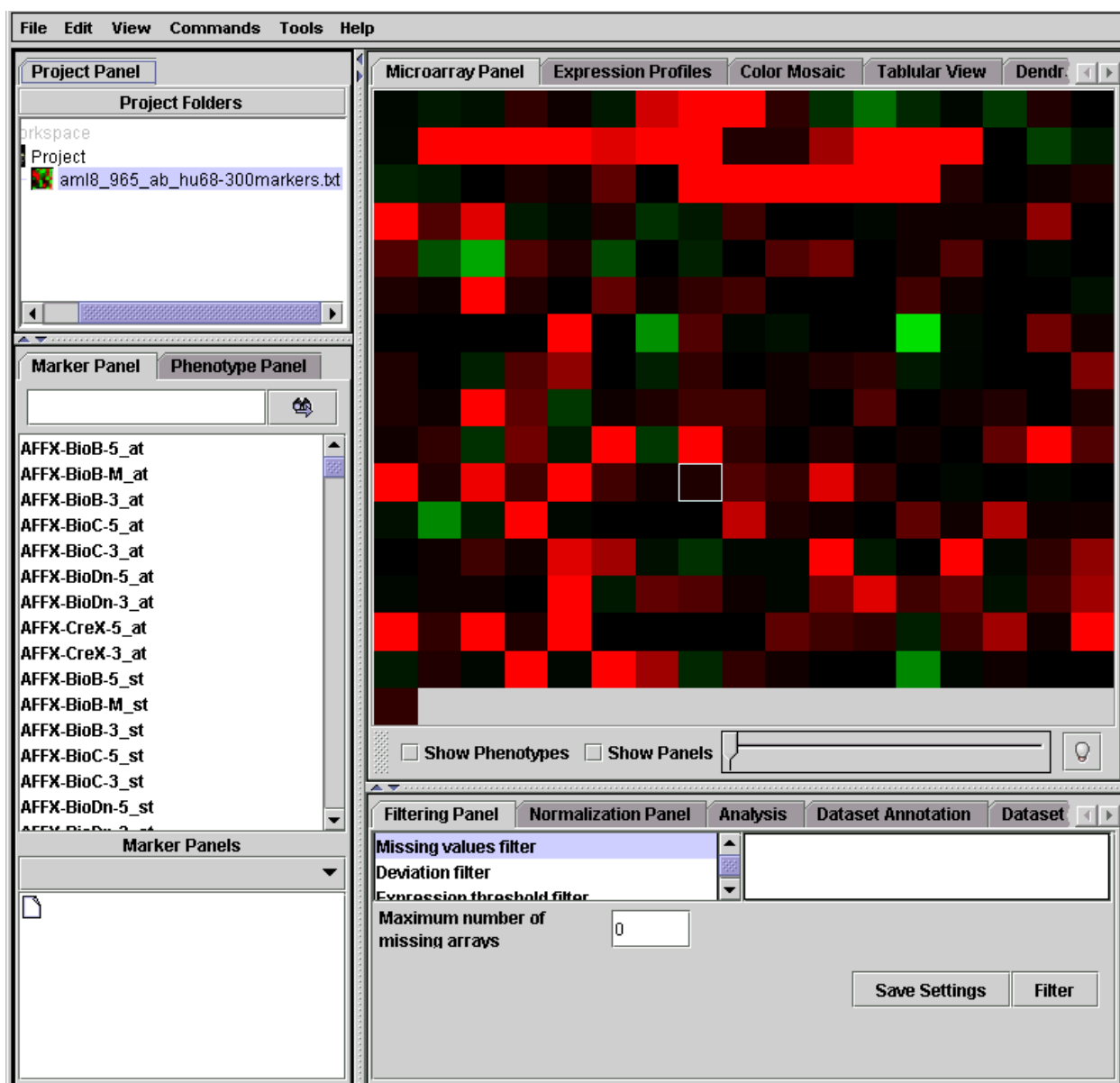


Figure 5.2-9 caWorkbench Display Immediately After Loading a Data File

Gene probes can appear in both the master list of probes as well as in the individual panels. In the former case, the pop-up options include adding that probe to a panel, clearing the selection of the probe, and viewing the annotation for the probe in the View window.

Panel sets have a number of associated options; several manipulate the individual panels contained therein. One option is to simultaneously *activate* all of the marker panels in a particular set. As described in the next section, most of the visualization tools in the View window provide a **Show Panels** checkbox.

When this option is selected, the tool's display will capture only that information contained in activated panels. A panel set's **Activate All** option can be used to activate all of the panels in that

set. Similarly, the **Deactivate All** option will deactivate all panels in the set. New panels can also be explicitly added to a set, using the **New Panel** option.

As mentioned in the previous section, it is also possible to save and load panel sets independent of the workspace where they were created. While it does not make sense to load a panel set generated from a data set that is not currently loaded, this facility can be useful when several saved workspaces share common data. The **Load Panel Set** option allows users to load panel sets defined outside the current workspace.

Since panel sets are automatically named consecutively as “New Set,” “New Set1,” “New Set2,” etc., it is highly possible that the name of a newly loaded panel set may conflict with other existing panel sets in the current workspace. In this case, caWorkbench automatically renames the newly loaded panel set with the next available consecutively numbered name. The user has the option of assigning more meaningful names however, using the **Rename Panel Set** operation.

Finally, individual marker panels can also be explicitly renamed, activated, deactivated and deleted from the panel set, and individual gene probes can be deleted from a panel. All of the elements listed in the Marker/Phenotype windows have pop-up menus associated with them, as indicated in the preceding figures. These pop-up menus are summarized in Figure 5.2-10.

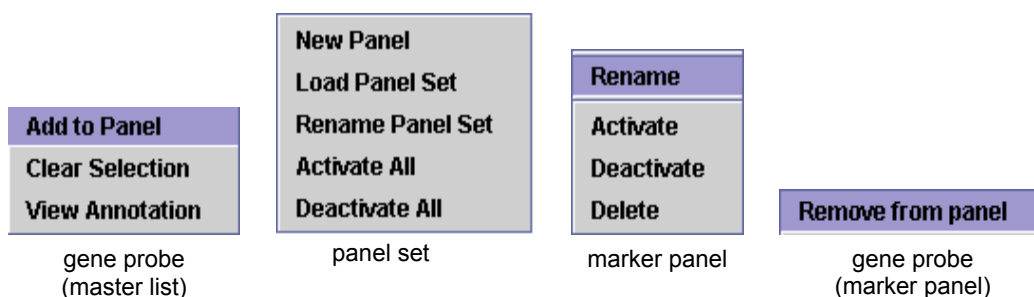


Figure 5.2-10 Pop-up Menus in the Marker Panel Windows

#### 5.2.4.2 Phenotype Panel Sets and Individual Phenotype Panels

caWorkbench uses the term *phenotype* to refer to any user-defined grouping of microarrays. These microarrays will often share some common property that in most cases is phenotypic, although this is not a requirement. For example, one such “phenotype” might be a single experiment on a tumor tissue sample, with a second “phenotype” defined as a collection of experiments performed on normal tissue samples.

Like the Marker Panel window, the Phenotype Panel window has two portions: the top portion lists the set of arrays included in the selected data set, and the bottom portion (titled “Phenotype”) lists any user-defined array groupings. These groupings are called “phenotype panels” or “phenotypes”—the assumption being that, in most cases, all of the arrays within a phenotype panel will share the same phenotypic characteristic. The name of the panel should be chosen as a mnemonic for this common phenotypic characteristic.

For data sets involving a single array, only that array is present in the top portion. But in the case of merged data sets, the display becomes more interesting: each experiment that was included in the set is displayed as a potentially separate phenotype.

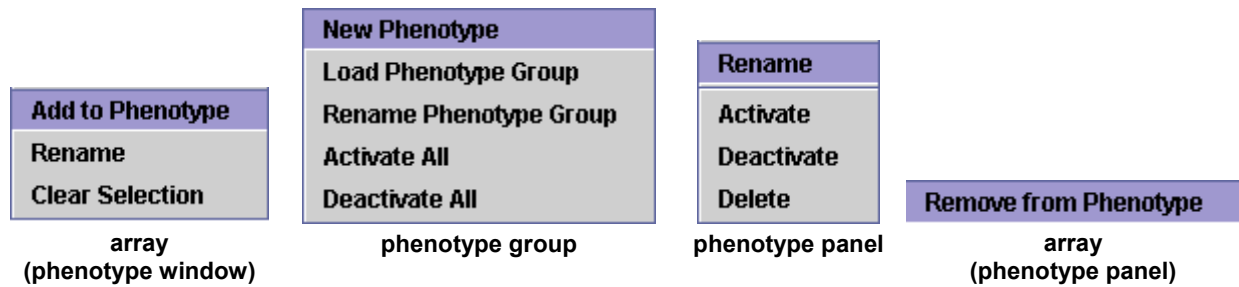


Figure 5.2-11 Pop-up Menus in the Phenotype Panel Windows

Analogous to the procedures for selecting markers into marker panels, arrays are selected and grouped—according to the user’s preferences—into phenotype panels. Each array in the top portion of the phenotype window has an associated pop-up menu with options **Add To Phenotype**, **Rename**, and **Clear Selection**. As in the case of working with marker panels, clicking on **Add To Phenotype** before creating a new phenotype panel (or selecting a previously defined one) will lead to the automatic creation of a new phenotype panel containing the selected array.

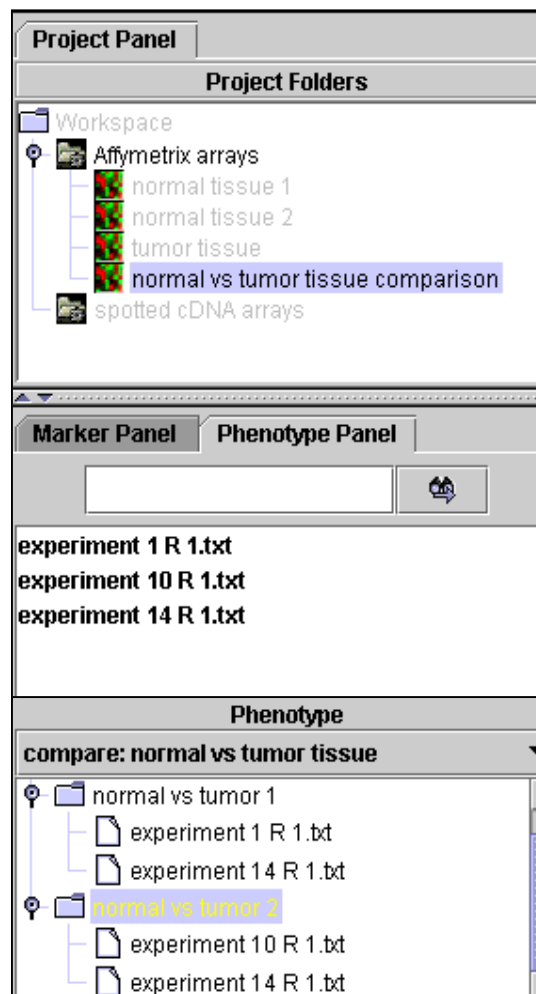


Figure 5.2-12 Defining Phenotype Panels

The pop-up menus associated with arrays, phenotype groups, phenotype panels, and the individual arrays contained in a panel are shown in Figure 5.2-11. Here, a phenotype group is analogous to a panel set. Like marker panel sets, a phenotype group can be saved and/or loaded independent of any specific workspace. And like marker panels, individual phenotype panels can be activated and deactivated. Analogous to the Show Panels checkbox in the View windows, many of the visualization tools include a Show Phenotypes checkbox that can be used to filter the data points.

Figure 5.2-12 shows an example study involving three experimental data sets, where two of the data sets are derived from normal tissue and one is derived from tumor tissue. After loading the data sets, the user has created a fourth merged data set named “normal vs tumor tissue comparison,” which combines all three of the original arrays. Although the original data sets have been aliased in the Project window as “normal tissue 1,” “normal tissue 2,” and “tumor tissue,” when the merged data set is selected and displayed in the Phenotype Panel, the original names of the three data sets—namely, experiments 1, 10 and 14, respectively—are exposed.

In order to now compare the tumor sample to each of the normal samples independently, two *phenotype* panels have also been created, named “normal vs tumor 1” and “normal vs tumor 2.” Individual experiments listed in the Phenotype Panel window are added to the newly created panels by right clicking on the experiments and using the pop-up menu’s **Add To Phenotype** option. By defining these phenotype panels, the user has now effectively created phenotype masks that allow each normal sample to be viewed independently alongside the tumor sample.

#### 5.2.4.3 The Commands Menu

Many of the commands for manipulating marker and phenotype panels are also available from the **Commands** menu option in the main menubar, as shown in Figure 5.2-13.



Figure 5.2-13 The Commands Menu Options

#### 5.2.5 The View Window

The View window is in a sense the main work area as it provides all of the visualization tools in caWorkbench. Folder tabs running across the top of the screen provide access to these tools, which are summarized in Table 5.2-2 and described in more detail below.

Some of the visualization tools are only applicable to data sets involving more than one array; others are enhanced by applying filters and/or normalizations to the data, and two of the tools are only applicable to clustered data—the Dendrogram and SOM Clusters tools. This section

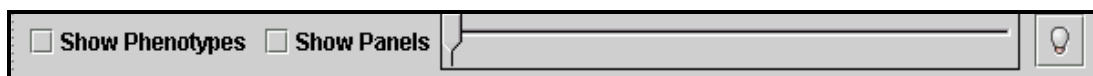
provides a quick tour of the general capabilities of those tools that can be applied to unclustered data. Information about applying filters and normalization in the context of these visualization tools is provided in [Section 5.2.6](#), along with a discussion of the Dendrogram and SOM Clusters tools.

**Table 5.2-2 Visualization Tools in the View Window**

<u>Visualization Tool</u>	<u>Description</u>
Microarray Panel	Displays expression measurements as spots over a red-green color spectrum.
Expression Profiles	Displays the expression of genes across several arrays/ hybridizations.
Color Mosaic	A color mosaic representation of measurements, with each array in one column and each probe in one row.
Tabular View	Presents the numerical values of the expression measurements in a table format; each row represents an individual probe and the columns display the signal and background intensities and intensity ratios.
Dendrogram	Displays tree-structured diagrams (dendrograms) reflecting the results of hierarchical clustering analysis.
SOM Clusters	Displays the results of self-organizing map cluster analysis.
caBIO Pathways	Displays BioCarta pathway diagrams for selected genes.
Marker Annotations	Allows users to retrieve and display CGAP annotations for genes within a marker panel.
Image Viewer	Displays snapshot images taken from whole screen views.

### 5.2.5.1 The Microarray Panel

The Microarray Panel is the default visualization tool in the View window, and is displayed when the application is first started. As each new data file is opened, that data set becomes the currently selected one, and the data is displayed in the Microarray Panel. The image displays color-coded levels of gene expression, using a color scale varying gradually from red (upregulated) to green (downregulated) through black (no change). The density of the data points in this screen is determined solely by the number of features on the array. The Microarray Panel has 4 controls at the bottom of the panel, shown in Figure 5.2-14.

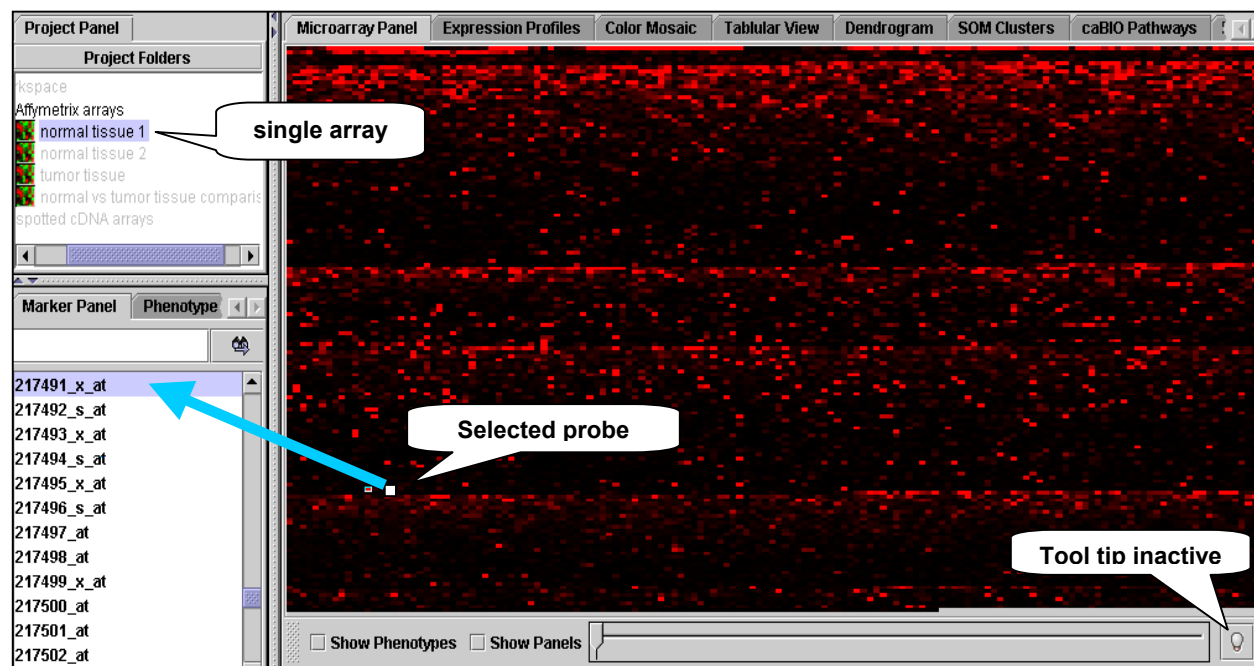


**Figure 5.2-14 Graphical Controls in the Microarray Panel**

The two checkboxes to the left, **Show Phenotypes** and **Show Panels**, determine which data points are included in the display. If neither is checked, then the entire data set is shown. The **Show Phenotypes** control is useful when working with data sets comprising multiple arrays. In this case, only those arrays that are included in a currently *activated* phenotype panel will be displayed (see previous section). Similarly, the **Show Panels** checkbox can be used to view only those probes that are currently included in an activated marker panel.

The scrollbar to the right of the two checkboxes is active only when a multi-array data set is being viewed. In this case the individual microarrays are displayed from left to right, and the scroll bar can be used to jump from one microarray to the next. The entry point to each of the individual chip displays is indicated by a tick mark on the scrollbar.

The light bulb icon to the right of the scrollbar activates a “mouse over tool tip” capability; when the light bulb is activated, mousing over and pausing on a square in the grid causes information about that probe to be displayed in a pop-up text box. In addition, if the current display is for a multi-array data set, mousing over the scrollbar pointer pops up a textbox identifying the array currently being viewed.



**Figure 5.2-15 Sample Display in the Microarray Panel**

The information that is displayed about individual probes when the mouse over capability is activated includes the probe set name, the expression value, a p-value designating the significance of the measurement, and a presence/absence designator, indicating whether or not the array includes a measurement for the probe. The p-value designation however, is only available if the underlying technology (e.g. Affymetrix platforms) supports it.

Figure 5.2-15 shows the display of a single Affymetrix array with over 22,000 features in the Microarray Panel. Neither the scrollbar nor mouse over tool tip is active in this example, as the display does not involve a merged data set and the light bulb icon is not illuminated.

The Microarray Panel provides an overview of the chip(s) under investigation and can be used for ascertaining the quality of the data—i.e., the uniformity of the hybridizations, the compatibility of intensities between chips, and so forth. Each feature on the chip can be accessed with a small cursor box and highlighted. Left-clicking the mouse will then highlight the corresponding probe in the master list contained in the Marker Panel window. This association between the Microarray and Marker Panel windows facilitates the selection of individual probes



for inclusion on explicit marker panels, as the user can then right click the selected probe and simply select **Add to Panel**.

Only those marker panels that are currently activated will be displayed when the **Show Panels** checkbox is checked. As described in [Section 5.2.4](#), marker panels can be activated and deactivated by right clicking on the panel itself, by activating and deactivating all panels within a marker panel set, or by using the Commands menu from the main menubar. Any number of smaller predefined marker panels can be activated and displayed in this zoomed view.

As described in Section 5.2.4, phenotype panels are used to create experiment groups. For instance, in a multi-array data set containing arrays from both normal and tumor tissue, the user may want to do pairwise comparisons of normal versus tumor samples. Alternatively, it may be interesting to segregate these samples into normal and tumor groups. Like their marker panel counterparts, phenotype panels can be selectively activated and displayed using the controls described above.

### 5.2.5.2 The Expression Profiles Tool

The Expression Profile view makes it possible to visualize changes in the gene expression levels across different hybridizations. This is useful especially in the analysis of time course or dose response experiments. Since the tool generates a graphical representation of *relative* expression levels across two or more arrays, the Expression Profile view can only be applied to merged and normalized data sets (see [Section 5.2.6](#)).

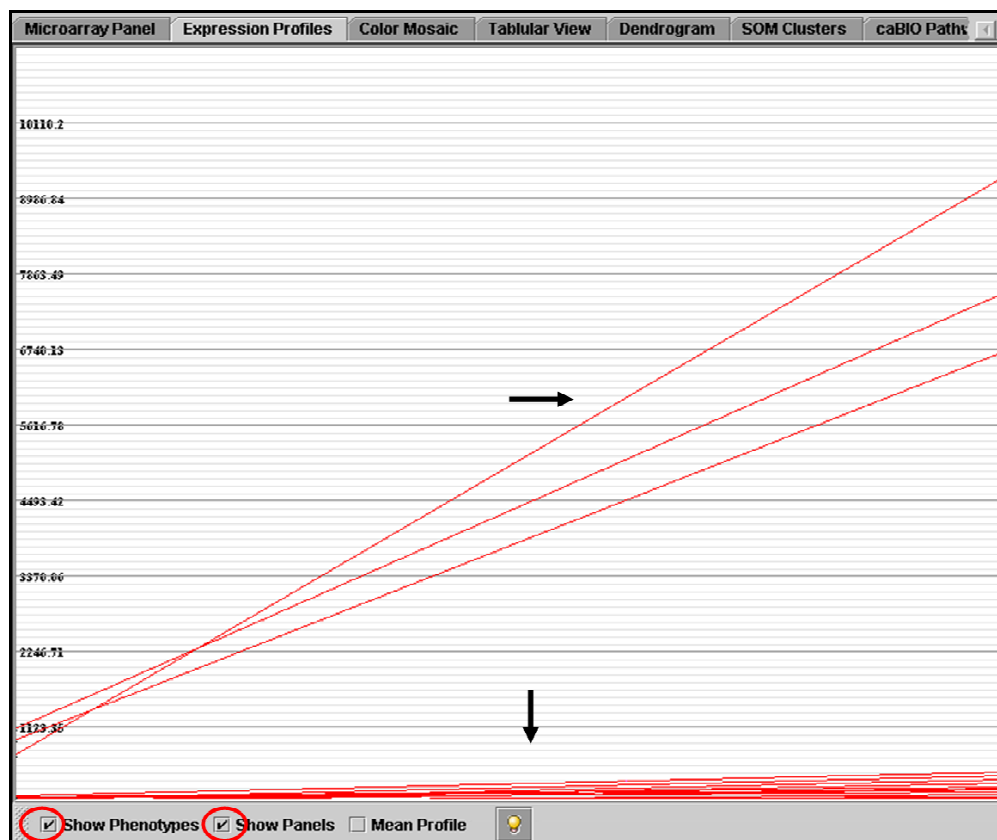


Figure 5.2-16 An Expression Profile Using Phenotype And Marker Panels

After loading, normalizing, and merging the desired data sets, the user may also wish to apply marker and phenotype panels in order to zoom in on the expression behavior of a subset of genes and/or hybridizations. Like the Microarray Panel display, the Expression Profile tool provides a mouse over tool tip capability that can be activated by clicking on the light bulb icon. The identities of the individual genes associated with the expression levels can then be obtained by mousing over the points of interest.

Figure 5.2-16 shows an example of an expression profile obtained from the study described in Section 5.2.4, where two of the data sets were obtained on normal samples and one data set was obtained on a tumor sample. These data sets were then merged to form a single data set containing three arrays, and the arrays were then grouped into two different phenotype panels, “normal vs tumor 1,” and “normal vs tumor 2.”

The expression profile in Figure 5.2-16 uses these phenotype panels, along with selected marker panels to focus on two disparate sets of genes. In one group the expression levels increase between phenotypes, whereas in the second group the expression levels are relatively unchanged between the two arrays.

### 5.2.5.3 The Color Mosaic View

When a Color Mosaic is applied to a merged data set of two or more microarrays, the gene expression levels across all of the microarrays are displayed as a color coded image, where each column in the image corresponds to one of the microarrays, and each row corresponds to a particular gene probe. Figure 5.2-17 shows a Color Mosaic generated for a merged data set containing two Affy chips.

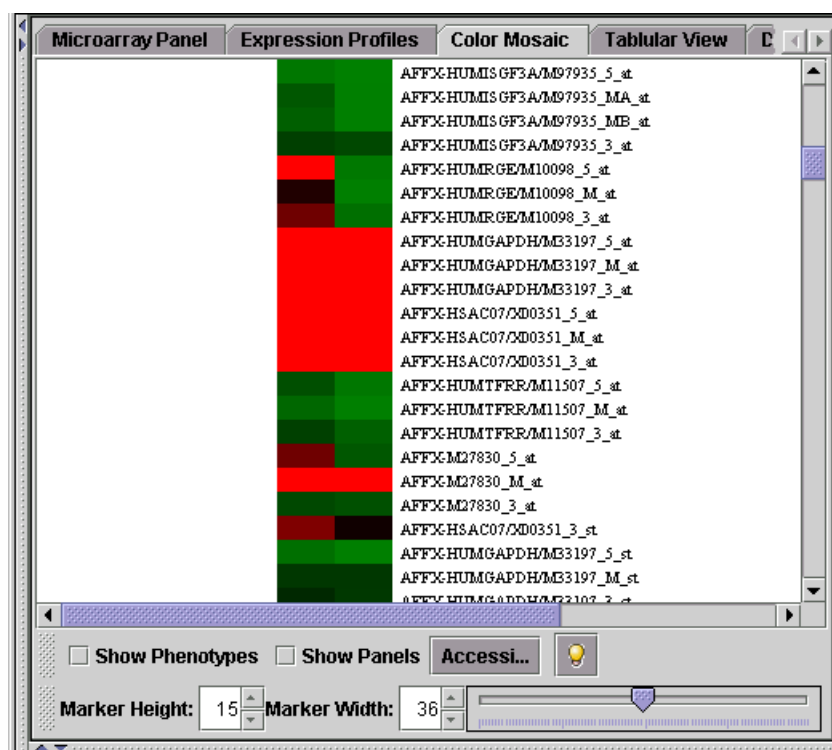


Figure 5.2-17 A Color Mosaic View of a Merged Data Set

The graphical controls provided in the Color Mosaic window are shown in Figure 5.2-17, and include checkboxes for selecting phenotype and marker panels, an *accession* button, a mouse over tool tip, controls for changing the height and width of the displayed markers, and a slider for modifying the intensity of the color codings.

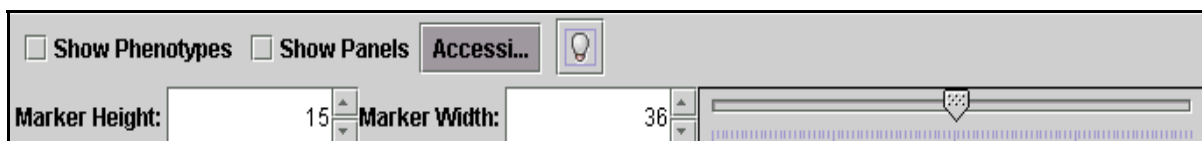


Figure 5.2-18 Graphical Controls in the Color Mosaic View

The accession button toggles on and off the display of the gene probe names for the individual markers. The tool tip, when activated, exposes the numerical intensity value when the mouse is poised over a specific tile. The height and width controls increase or decrease the dimensions of the tiles as well as the associated labels when these are displayed. The slider increases or decreases the thresholds used to defined the red/black/green color codings. Clicking on a tile in the mosaic will highlight the corresponding gene probe in the Marker Panel window, where it can be picked and added to a marker panel if desired.

#### 5.2.5.4 The Tabular View

Like the Color Mosaic, the Tabular View can be used to obtain a side-by-side comparison of the observed intensities for each gene probe over multiple chips. The display in this case however, shows the numerical values in a simple table format.

Expression Profiles					
Color Mosaic					
Tabular View					
Dendrogram					
SC					
Marker	Signal 532	Background...	Signal 635	Background...	Demo gpix..
IMAGE:390307	2166	42	1940	101	1.155
IMAGE:386559	2574	42	2465	100	1.071
IMAGE:390236	1123	40	1506	99	0.77
IMAGE:387243	2335	41	2052	102	1.176
IMAGE:374756	1411	41	1403	104	1.055
IMAGE:385380	1608	41	2299	102	0.713
IMAGE:388244	502	41	1093	102	0.465
IMAGE:406947	1443	41	1662	93	0.894
IMAGE:407081	1147	41	1008	92	1.207
IMAGE:407552	1893	41	1688	94	1.162
IMAGE:419771	2168	47	1942	100	1.151
IMAGE:408379	1431	47	1615	102	0.915
IMAGE:408542	1477	51	1368	101	1.125
IMAGE:407068	3762	42	4852	100	0.783
IMAGE:407202	1906	43	1977	97	0.991
IMAGE:407556	3349	49	4291	103	0.788
IMAGE:407924	2546	42	2229	97	1.174
IMAGE:408189	759	43	1315	98	0.588
IMAGE:420149	6392	41	3749	98	1.74
IMAGE:406745	2745	53	2349	119	1.207
IMAGE:419243	617	43	825	99	0.791
IMAGE:407558	1514	41	1531	97	1.027
IMAGE:408111	1815	42	1868	100	1.050

Figure 5.2-19 The Tabular View

As with most of the other panels in the View window, checkboxes are provided for displaying selected phenotype and marker panels. Two additional controls, **Signal** and **Background**, are available when working with GenePix data sets. Figure 5.2-19 shows a Tabular View generated for the *Demo gpix 1.gpr* data set included with the distribution in the *example data* directory.

For Affymetrix data sets, the **Signal** and **Background** options are not active. In both cases, any values that may have been filtered out by applying one of the masks described in [Section 5.2.6](#) will be highlighted in yellow in the column showing the probe-specific intensities.

### 5.2.5.5 The Marker Annotations and caBIO Pathways Views

The Marker Annotations View window is used to view and retrieve CGAP annotations for selected genes. Individual annotations can be viewed by right clicking on the selected gene in the master list and clicking the **View Annotation** option. Alternatively, a collection of annotations for a set of genes can be viewed when those genes are included in a marker panel. In this case, the annotations can be retrieved by activating the marker panel, and checking the **Use Panels** checkbox in the Marker Annotations View window.

Color Mosaic	Tabular View	Dendrogram	SOM Clusters	caBIO Pathways	Marker Annotations	Image Viewer
<a href="#">hypothetical protein FLJ10979</a>						
<a href="#">general transcription factor IIIC, polypeptide 1, alpha 220kDa</a>					<a href="#">mapo13Pathway</a>	
<a href="#">comparative gene identification transcript 94</a>						
<a href="#">Homo sapiens transcribed sequences</a>						
<a href="#">small proline-rich protein 2C</a>						
<a href="#">ubiquitin specific protease 14 (tRNA-guanine transglycosylase)</a>						
<a href="#">Homo sapiens transcribed sequences</a>						
<a href="#">chromosome 17 open reading frame 28</a>						
<a href="#">zinc finger protein 495</a>						
<a href="#">Homo sapiens transcribed sequence with weak similarity to protein refNP_055301.1 (H.sapiens) neuronal thread protein [Homo sapiens]</a>						
<a href="#">hypothetical protein FLJ33071</a>						
<a href="#">interleukin 12B (natural killer cell stimulatory factor 2, cytotoxic lymphocyte maturation factor 2, p40)</a>					<a href="#">cytokinePathway</a>	
					<a href="#">th1th2Pathway</a>	
<a href="#">TAR DNA binding protein</a>						
<a href="#">Homo sapiens mRNA; cDNA DKFZp586O0724 (from clone DKFZp586O0724)</a>						
<input checked="" type="checkbox"/> Use Panels	<input checked="" type="checkbox"/> Show Pathways			<input checked="" type="checkbox"/> Sort By Pathway		

Figure 5.2-20 The Marker Annotations View

Figure 5.2-20 shows a part of the Marker Annotations view for a selected panel. The first column lists the name of the gene, and, when the **Show Pathways** checkbox is checked, the second column lists any pathways associated with that gene.

The third option, **Sort by Pathway**, will be available in the caWorkbench v2.0 release. Clicking on the gene name will open a new browser window displaying the CGAP annotation page for that gene. Clicking on a pathway will make the caBIO Pathways window active in the View window, and display the corresponding BioCarta pathway. Figure 5.2-21 shows one such pathway diagram.

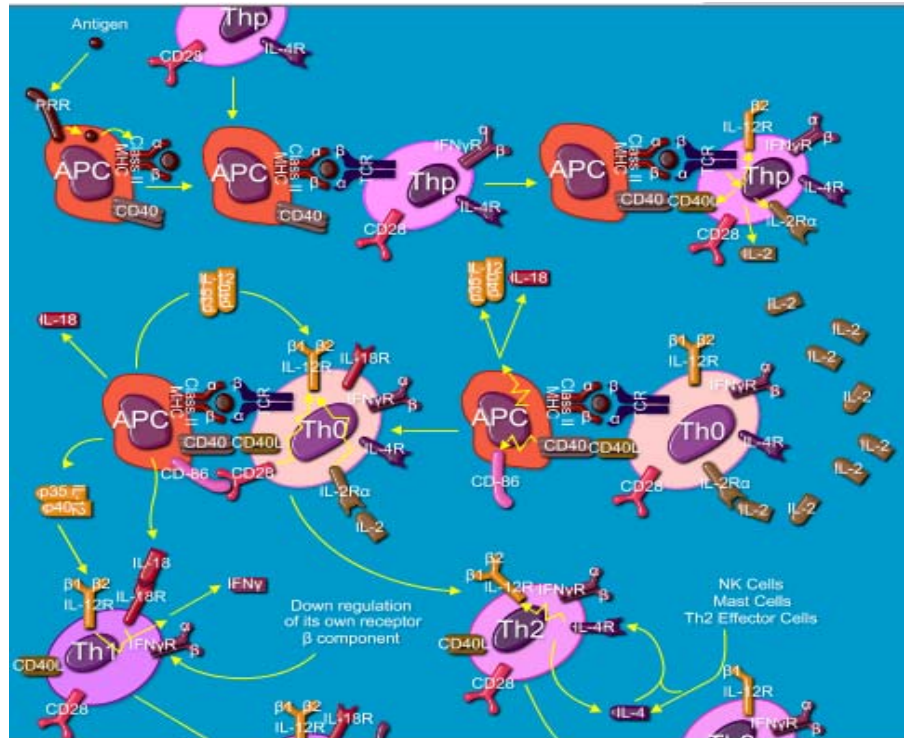


Figure 5.2-21 The caBIO Pathways View

### 5.2.5.6 The Image Viewer

Several of the visualization tools provide a means of capturing a snapshot of a selected region of the display. For example, right clicking on any point in the Microarray Panel display will cause a pop-up text control to appear saying **Image Snapshot**. Left-clicking on this control will create a snapshot of whatever is currently visible in the display view, and store that image under the associated data file in the project window. Figure 5.2-22 shows a snapshot captured from a portion of a merged data set.

Note that the image file in the project window is named “Microarray Image: Affy1.txt,” but is stored under the merged array data—not under the *Affy1.txt* data set itself. This is because the snapshot image is always stored with whatever data set was active in the view window where the snapshot was taken. Because merged data sets maintain the segregation of the individual phenotypes, the stored image name reflects that data set from which those data points were taken. Moreover, because the display was much smaller when the shot was captured, the image fills only a part of the Image Viewer window. An image can be saved in TIFF, JPEG, or PNG format by selecting the image in the Project window, and selecting the **Export** option from the drop-down file menu.

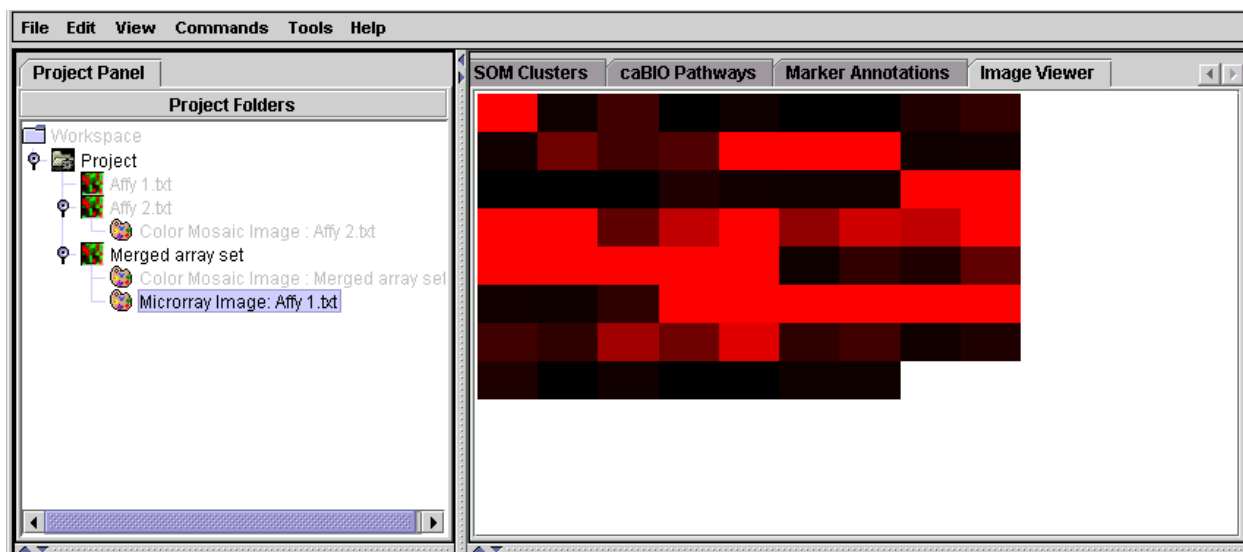


Figure 5.2-22 Using The Image Viewer

The steps used to capture an image and to view and save the stored snapshot are:

1. In the Project window, select the data set for which you would like to capture an image.
2. In the View window, select a visualization tool.
3. Right click on any point in the tool's display and left click on **Image Snapshot**.
4. In the Project window, expand the associated data set (if is not already open) by clicking on the key-like icon to the left of the data set.
5. Double click on the stored image to bring it up in the Image Viewer.
6. Select the image in the Project window, and use the **Export** option from the file menu to save it in TIFF, JPEG, or PNG format.

### 5.2.6 The Analysis/Annotation Window

The tab-indexed tools in this last window include facilities for filtering, normalizing, and analyzing data, along with panels for viewing the history of operations that have been performed on a data set and general experiment and annotation information. All of the filtering, normalization, and analysis tools include a **Save Settings** option, which saves the parameters used in the analysis or processing step with the workspace.

Filters are used to remove data points when some quality data criteria are not met. As a result of applying a filter, the status call of a questionable data point may be reset to "missing," or alternatively, the data point may be eliminated altogether from the data set. In the later case, all measurements for that marker (across all chips in the data set being filtered) will be eliminated. In contrast, normalizers do not change the status or remove individual markers, but re-scale the observed intensities, usually in preparation for some type of analysis. Filtering or normalizing a dataset *A* will give rise to a new dataset *B*, with *B* appearing as a child of *A* in the project tree window.

### 5.2.6.1 Filtering Operations

The Filtering Panel contains several filters that allow the software to set certain values to missing. For instance, the *Affy detection call* filter allows the user to filter out undesirable values on the basis of the Affymetrix calls (“present,” “absent,” or “missing”). Figure 5.2-23 shows the result of applying this filter to set all markers with a call of “absent” to “missing.” In the Microarray Panel display of the filtered data, all of the missing data points now appear as grey squares in the heat map. Table 5.2-3 summarizes the filters that are available from a pull-down list in the Filtering Panel window.

Table 5.2-3 The Filtering Panel Toolset

<u>Filter</u>	<u>Description</u>
Missing values	Discards all markers that have “missing” measurements in at least $n$ microarrays, where $n$ is defined by the user. Another filter must first be applied however, in order to generate the missing values upon which this filter can operate.
Deviation	Sets all markers whose measurements deviate below a given value across all microarrays as missing.
Expression threshold	Sets all markers whose measurements are inside (or outside) a user-defined range as missing.
Affy detection call	Applicable to Affymetrix data only. Sets all measurements whose detection status is any user-defined combination of A, P or M as missing.
2-channel threshold	Applicable to 2-channel arrays (Genepix) data only. Defines applicable ranges for each channel, and sets all values for which either channel intensity is inside (or outside) the defined range as missing.

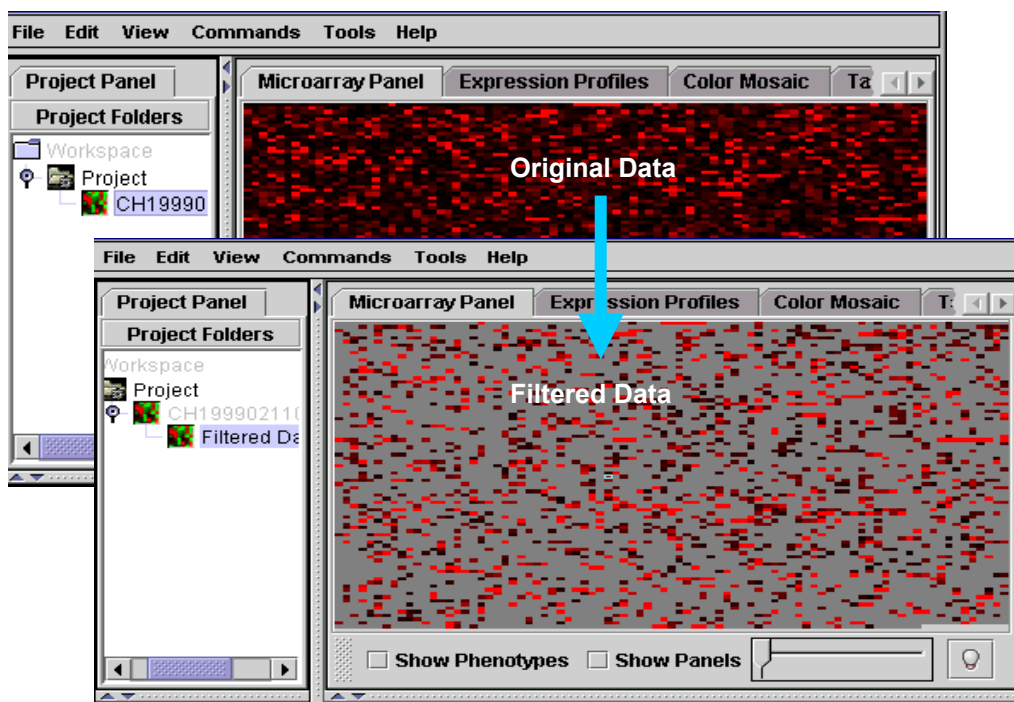


Figure 5.2-23 Using the Affy Detection Calls as a Filter



### 5.2.6.2 Normalization Tools

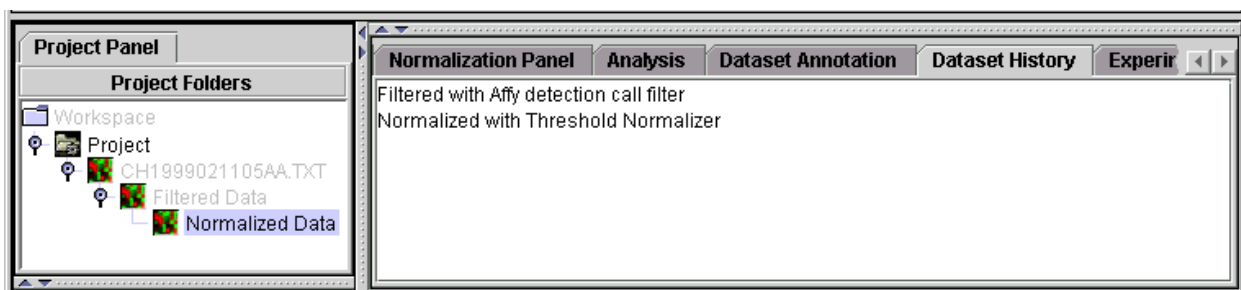
Before comparing multiple microarrays to one another, the user must first ensure that the observed values therein have been made “comparable” through a process of normalization. The normalization panel offers the user several gene-centric or array (tissue)-centric methods that are summarized in Table 5.2-4.

**Table 5.2-4 The Normalization Panel Toolset**

<b>Normalization Tool</b>	<b>Description</b>
Missing value calculation	Replaces every missing value with either the mean value of that marker across all microarrays, or with the mean measurement of all markers in the microarray where the missing value is observed.
Log2 Transformation	Applies a log <sub>2</sub> transformation to all measurements in a microarray.
Threshold Normalizer	All data points whose value is less than (or greater than) a user-specified minimum (maximum) value are raised (reduced) to that minimum (maximum) value
Marker-based centering	Subtracts the mean (median) measurement of a marker profile from every measurement in the profile
Array-based centering	Subtracts the mean (median) measurement of a microarray from every measurement in that microarray.
Mean-variance normalizer	For every marker profile, the mean measurement of the entire profile is subtracted from each measurement in the profile and the resulting value is divided by the standard deviation.

### 5.2.6.3 The Dataset History window

caWorkbench provides a convenient system for electronic tracking of all actions. As noted, each time a new file or image is generated, that file appears in the Project Tree Window as a new node occurring beneath the data set from which it was derived. In addition, the Dataset History window displays a list of all of the operations that were performed on both the currently selected data set as well as on all of its “parent” data sets in the Project Tree. Figure 5.2-24 shows the history window for the normalized data shown in the previous figure.



**Figure 5.2-24 The Dataset History Window**

### 5.2.6.4 The Analysis Tools

caWorkbench provides two clustering methods in the Analysis panel. Hierarchical clustering groups markers into clusters on the basis of similarities in their expression profiles to form a hierarchical tree that can be viewed in the Dendrogram View window (see Figure 5.2-25).

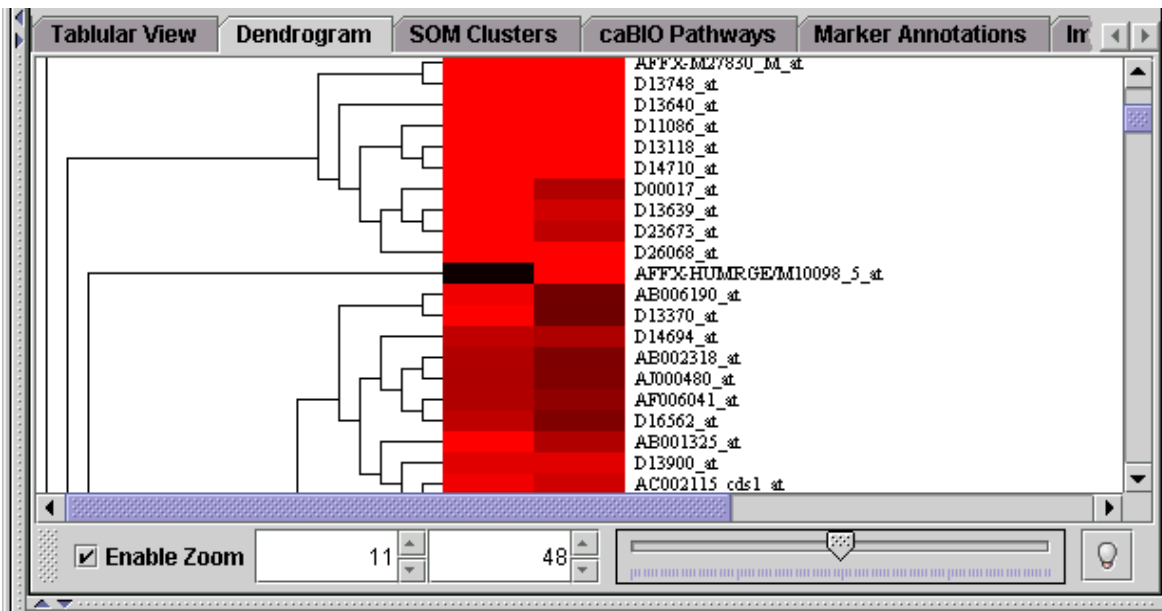


Figure 5.2-25 An Example of Hierarchical Clustering

SOM analysis uses self-organizing neural nets to identify genes with similar expression patterns, and maps expression profiles into the cells of user defined grids. The SOM Clusters View can then be used to explore the resulting maps. Figure 5.2-26 shows the same data set as in Figure 5.2-25 in the SOM Clusters View window.

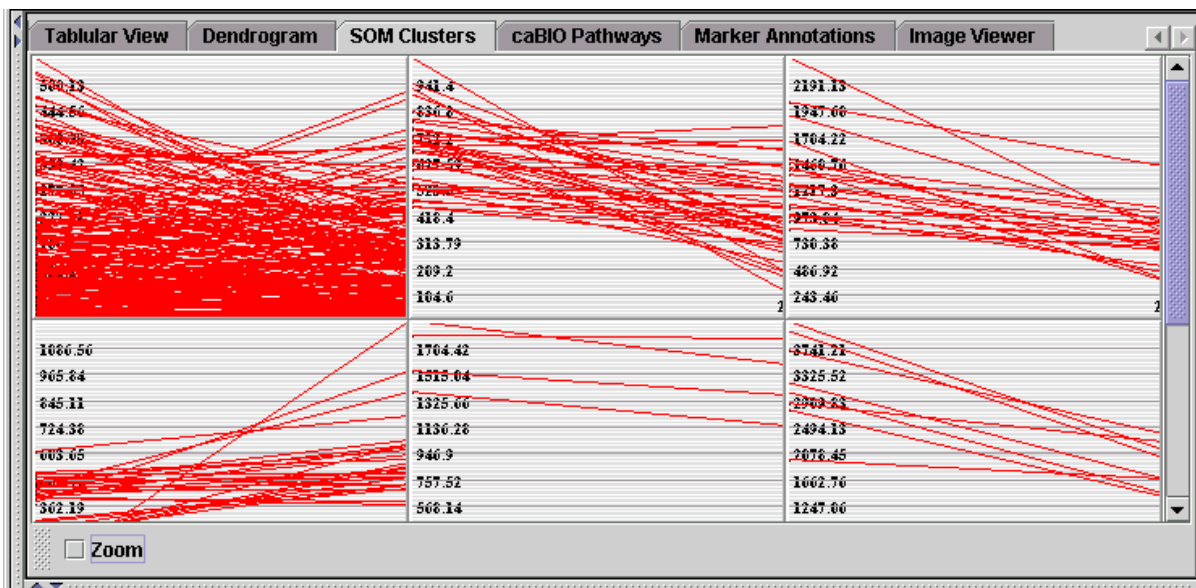


Figure 5.2-26 The SOM Clusters View window

caWorkbench's implementation of these algorithms is based on their implementation in the [Multi Experiment Viewer \(MEV\)](#) platform, which is freely available from The Institute for Genomic Research (TIGR).

#### **5.2.6.5 The Dataset Annotation Tool**

This panel provides a simple text window for adding any textual information that the user wishes to associate with a particular data set. Examples might include annotations found on the CGAP web site, questions that arise during the analysis which the user may wish to pursue at a later time, actions that were taken that are not otherwise tracked by caWorkbench, etc.

Cut and paste operations are supported to facilitate importing and/or exporting text to and from this window. In particular, as the “parent” data set’s annotations are not inherited by the “child” nodes, the user may wish to copy and paste some of these as new data sets are derived. Any text entered in this window will be saved and retrieved with the experiment when the workspace is reopened.

#### **5.2.6.6 The Experiment Info Tool**

This read-only text window displays the textual preamble that precedes the data in most experiments. While it is not possible to modify the text in this display, the user can copy that text if desired into a Dataset Annotation panel. For example, when two independent data sets are merged to form a new data set, the latter has no experiment information associated with it. Using copy and paste operations, the user can copy the experiment information from each of the original data sets into the Dataset Annotation window for the merged data.

### 5.3 The Comparative Genomic Hybridization Viewer: webCGH

Microarray-based Comparative Genomic Hybridization (CGH) is a powerful tool for high throughput screening of DNA copy number changes on a genome-wide scale. In array-based CGH, thousands of genomic BAC<sup>25</sup> or cDNA clones are arrayed on a microscope slide. When equal amounts of normal and tumor DNA—labeled with different colors—are then hybridized on the array, the resulting differences in the observed signal intensities indicate differences in the copy number (i.e., chromosomal aberrations) for that particular area of the genome.

To facilitate the discovery and analysis of genes included in these chromosomal aberrations, NCICB has developed a CGH viewer that allows the user to visualize copy number changes in the genome, and to identify the genes that are located in the rearranged chromosomal regions.

webCGH is a web-based application for visualizing and mining microarray-based CGH data. webCGH allows users to search the CGH experiments database; to create persistent user-defined groupings of experimental bioassays; and to generate whole genome and annotation plots with zoom capabilities for focusing on chromosomal regions of interest.

To start using webCGH, point your browser to the following URL:

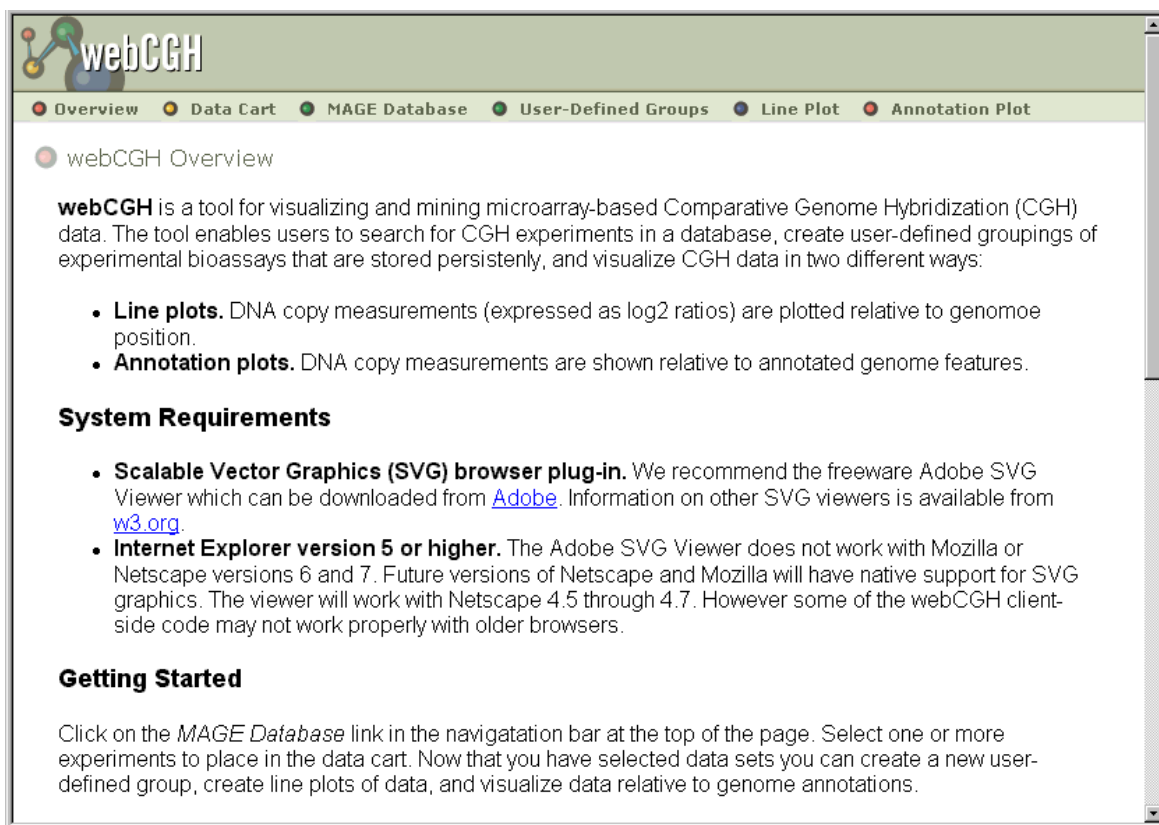


Figure 5.3-1 webCGH Overview

Figure 5.3-1 shows the webCGH Overview page that is displayed when you first enter the web site. This overview summarizes the tools available for analysis, the system requirements and

<sup>25</sup> Bacterial Artificial Chromosome

recommended software for using these tools, and tips for getting started. The navigation bar appearing at the top of each webCGH page contains links to the experimental database, to the user's "data cart," to the available tools, and to this overview page itself.

### 5.3.1 Selecting Experiments for Analysis

The first step in using the CGH viewer is to select the data that you would like to analyze from the experimental database. Clicking on the [MAGE Database](#) link in the navigation bar takes you to a page listing the available experiments, as shown in Figure 5.3-2. To select experiments for analysis, click on the checkboxes associated with those data and press the [Add to Cart](#) key.

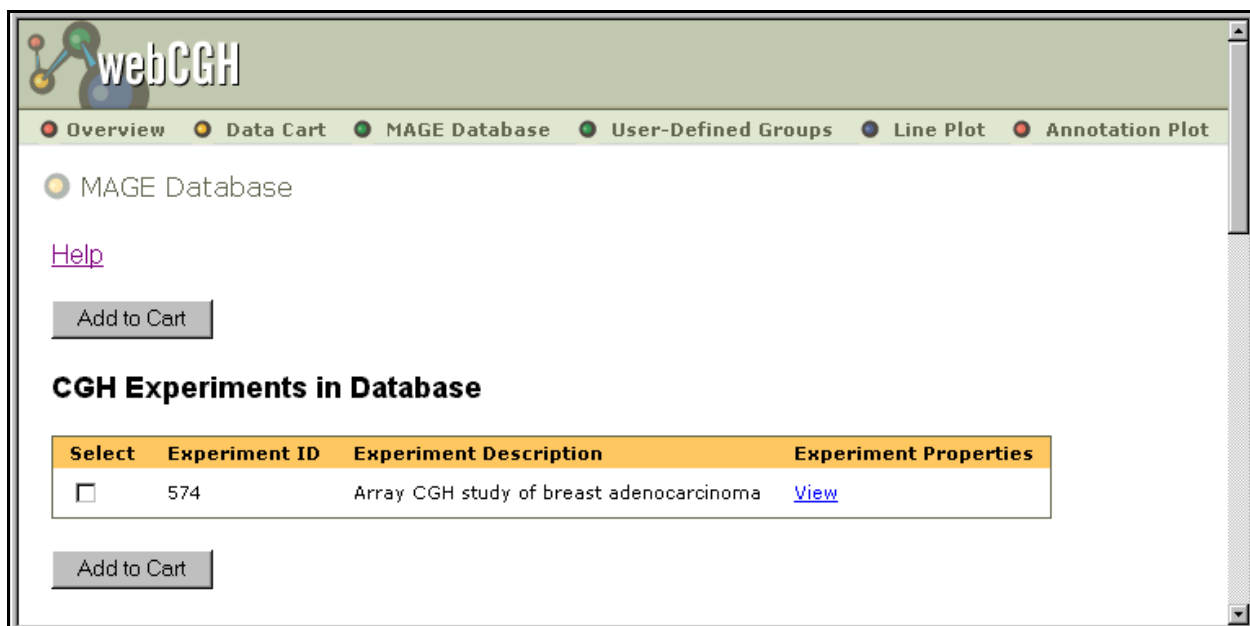


Figure 5.3-2 Selecting Experiments for Analysis

Each entry in the table in Figure 5.3-2 lists the database identifier (Experiment ID) and a brief description of the experiment, along with a hot link ([View](#)) to the experiment description and a list of all the bioassays (hybridizations) available for that experiment

Clicking on [Add to Cart](#) causes all of the selected experiments to be added to the user's data cart, and the screen is refreshed to show the current contents of the cart. The cart display page looks very much like the MAGE Database page. A similar table format is shown, but the title of the page is changed to "Data Cart," and the [Add to Cart](#) key is changed to [Remove](#). The Data Cart page can be accessed at any time by selecting [Data Cart](#) from the navigation bar.

With the exception of the Overview page, each of the webCGH pages provides a simple context-sensitive help tool. Clicking on the [Help](#) key from any screen will display a pop-up window displaying helpful information and tips for interacting with that screen.

### 5.3.2 Working with User-Defined Groups

User-defined groups allow researchers to create, display, and delete bioassays according to his or her preferences. This is useful when several experiments are included in the cart, and the user wants to create new subgroups or to combine bioassays from different experiments for analysis.

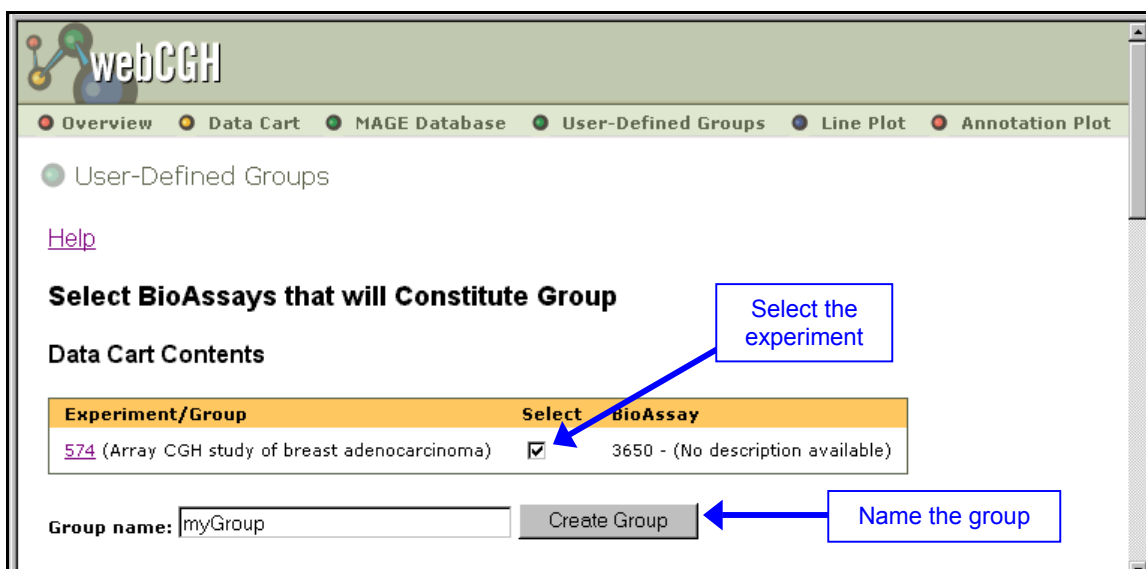


Figure 5.3-3 Defining Experiment Groups

Figure 5.3-3 shows the screen for defining experimental groups. Only those experiments that have previously been transferred to the cart are available for inclusion in new experimental groups. Each bioassay to be included in the group is selected by clicking on its associated checkbox. After selecting all of the bioassays to include in the new group, the user enters a name in the **Group name** textbox, and presses the **Create Group** button to create the group.

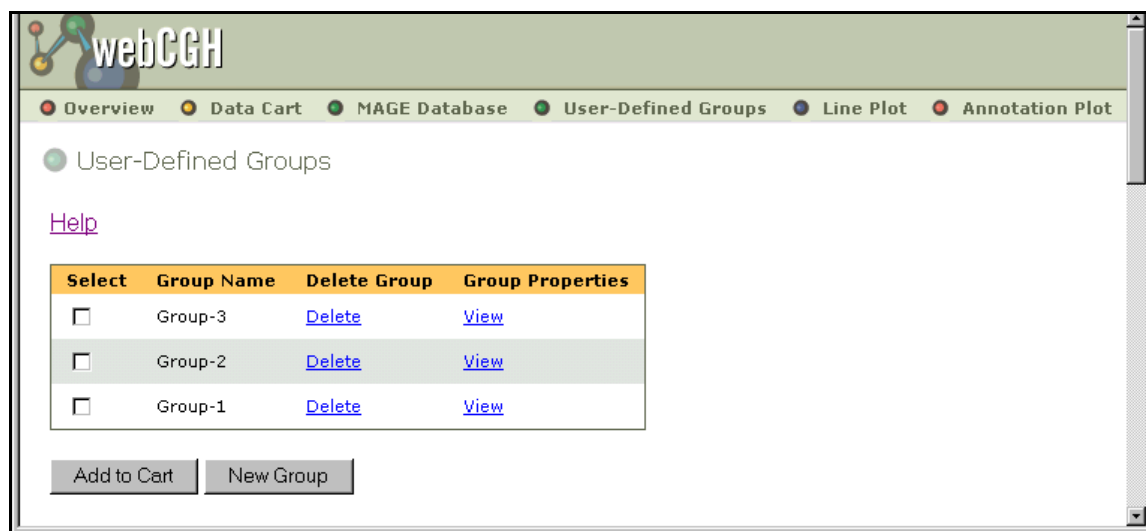


Figure 5.3-4 Summary of User-Defined Groups

All user-defined groups are stored persistently with the user's browser as cookies. Group names can use any combination of alphanumeric characters, along with the special characters '-' and '\_'. Defining a new group with a group name that has been used previously will lead to a replacement of that previous definition.

Each time a new group has been created, a summary page listing the currently defined groups is displayed, providing the user with further options for viewing or modifying the groups. As shown in Figure 5.3-4, the summary page also allows the user to add selected groups to the cart.

### 5.3.3 Working with Line Plots

The line plot provides a summary of DNA copy number measurements across the whole genome, with log 2 ratios of the observed probe intensities (sample versus control) calibrated on the Y axis and genome location along the X axis. The line plot requires prior loading of data into the data cart.

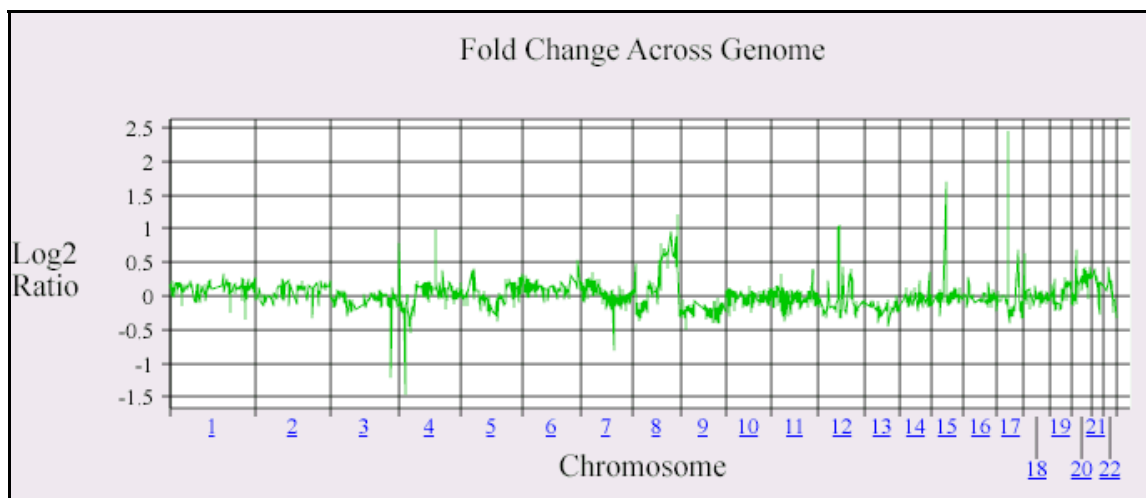
**Figure 5.3-5 Generating a Line Plot**

As shown in Figure 5.3-5, the Line Plot Tool provides several options for plotting the data. The user may plot data from each array individually or plot mean values for the group, by selecting the appropriate radio button, **Plot individual arrays** or **Plot group/experiment averages**, respectively. In the latter case, a mean of the log ratios with a 95% confidence interval will be plotted in the graph. Due to the density of data points in whole genome plots, the confidence intervals will only be displayed once the user has drilled down to a size less than or equal to a single chromosome. The user can also choose to display “raw” or “smoothed” data. Figure 5.3-6 shows an example line plot displaying raw data.

Because microarray-based CGH experiments measure large scale alterations to chromosomes rather than individual gene expression levels, a reasonable method to minimize non-biological sources of noise is to average the probe values from the same array within a “sliding window” based on genome position. The “window size”—defined by the user—defines the number of probes to be included in calculating these averaged values. Thus, with a window size of 10 (the default), the normalized value of the probe at position 5 in the window is the mean raw array value of all (10) probes in the window.



As this is a *sliding* window, each data point reflects a newly calculated average of the immediately surrounding raw values.<sup>26</sup> The user can also determine the size of the output window, using the controls shown in Figure 5.3-5 to set the width and height of the plots (in pixels) to any positive integer.



**Figure 5.3-6 Line Plot Display of Raw Data**

Additional display options become available once the plot is generated. When multiple experiments are displayed in a single plot, a table in the upper right-hand corner of the display (Figure 5.3-7) lists the experiments and provides radio buttons and checkboxes for suppressing/displaying or highlighting selected experiments.

Display	Highlight	Experiment	Bioassay
<input checked="" type="checkbox"/>	<input type="radio"/>	2959	5102
<input checked="" type="checkbox"/>	<input type="radio"/>	2960	5103

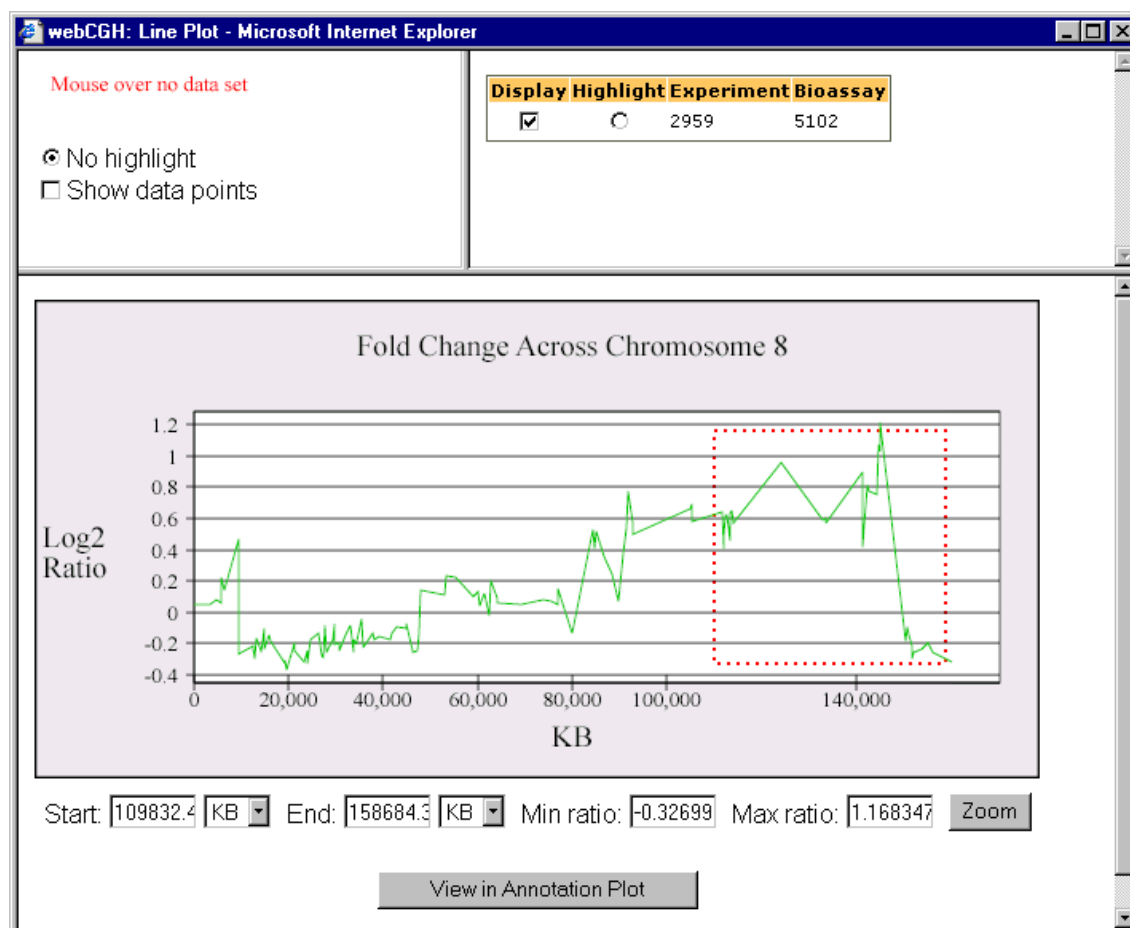
**Figure 5.3-7 Suppressing, Displaying, and Highlighting Selected Experiments**

Generally, each plotted microarray experiment has its own color, and mousing over a plot will reveal the name of the underlying experiment. If the user chooses to examine one plot at a time, he can deselect the other plots or determine which plots he wants to view together by checking the display box of any particular experiment. The experiments table lists the name of each experiment and indicates whether the values from a single micro array or the mean values of several micro arrays are displayed.

The chromosome number hyperlinks running along the X axis allow users to zoom in to the individual chromosomes. Clicking on a chromosome number opens a new browser window displaying a graph pertaining to the selected chromosome only. In Figure 5.3-8, chromosome 8 has been selected from the plot in Figure 5.3-6. The X axis now shows the distance between data points in kB. In addition, the graph can be shown with or without the data points underlying the

<sup>26</sup> Note that this normalization method only considers the order of probes and not their absolute position with regard to the chromosome. This is an obvious shortcoming.

analysis, by selecting (or deselecting) the **Show data points** checkbox. When the individual data points are displayed, the user can mouse over these measurements to reveal the clone name.



**Figure 5.3-8 Drilling Down to a Single Chromosome**

The researcher can continue to drill down to even smaller chromosomal regions by specifying the exact stretch of DNA (start and endpoint in kB) and a log ratio range (min ratio/max ratio). These values can be entered directly in the textboxes appearing at the bottom of the screen in Figure 5.3-8. Alternatively, the mouse pointer can be used to select a desired area directly from the plot, by left clicking and holding down the mouse button while dragging the mouse over the desired region. The selected area is then outlined in the graph (see Figure 5.3-8), and the distance and ratio boxes are automatically filled with the corresponding values. Clicking on the zoom button then opens a new window displaying the selected region.

Figure 5.3-9 shows the new graph defined by the area selected in Figure 5.3-8. Though not shown in Figure 5.3-9, the same controls for further drilling down to increasingly higher levels of resolution are present in each successive screen—only the graph itself changes.

Once a sufficient resolution for the region of interest has been achieved, the user can select the **View in Annotation Plot** button to proceed to a display showing the DNA copy number measurements graphically integrated with annotated genome features.

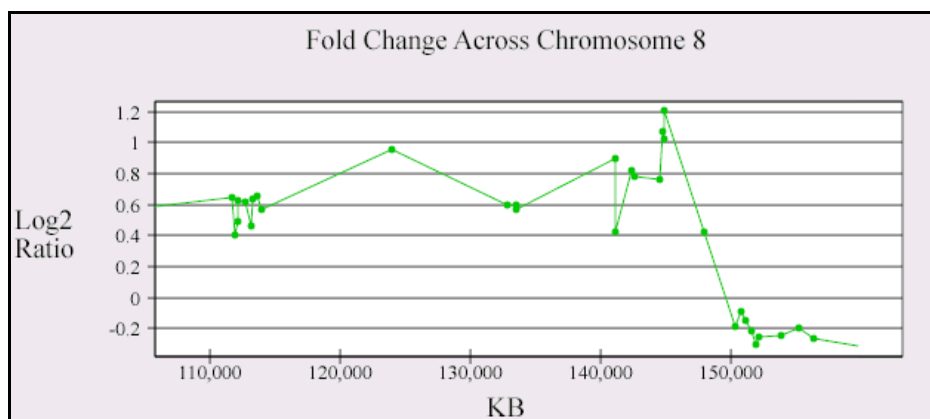


Figure 5.3-9 Zooming in on a Region of Interest

### 5.3.4 Working with Annotation Plots

The Annotation Plot tool can be accessed from either the webCGH navigation bar or directly from an existing line plot. Figure 5.3-10 shows the top half of the first screen associated with the tool. Opening the tool from the navigation bar requires that the user first specify the chromosome number and the region's end points—in kilobytes, megabytes, or base pairs. In contrast, when accessed from a given line plot, these values are automatically filled in.

Figure 5.3-10 The Annotation Plot Tool (top half)

Additional parameters similar to those required for generating a line plot must also be designated. Again, the options include plotting individual arrays or grouped averages, and viewing the raw or smoothed data points. Further options specify whether or not the names of the annotation features should appear in the plot itself, along with saturation values for color coding the observed copy measurements.

The upper and lower limits of saturation indicate the upper and lower limits that will be used to color code the probes in the annotation plot. All of the probes with copy number ratios equal to or less than the lower limit will be shown in blue, and all of the probes with copy number ratios equal to or greater than the upper limit will be shown in red.

Possible genome annotation features to include in the plot are presented in the lower half of the screen (Figure 5.3-11), where the user can choose those features he wants displayed.

### Genome Annotation Types

<input type="checkbox"/> <a href="#">acembly</a>	<input type="checkbox"/> <a href="#">affyGeno</a>	<input type="checkbox"/> <a href="#">affyRatio</a>	<input type="checkbox"/> <a href="#">all_bacends</a>
<input type="checkbox"/> <a href="#">all_fosends</a>	<input type="checkbox"/> <a href="#">all_sts_primer</a>	<input type="checkbox"/> <a href="#">all_sts_seq</a>	<input type="checkbox"/> <a href="#">bacEndPairs</a>
<input type="checkbox"/> <a href="#">blastzBestMm3</a>	<input type="checkbox"/> <a href="#">blastzTightMm3</a>	<input type="checkbox"/> <a href="#">blastzTightRn2</a>	<input type="checkbox"/> <a href="#">blatFuqu</a>
<input type="checkbox"/> <a href="#">est</a>	<input type="checkbox"/> <a href="#">gl</a>	<input type="checkbox"/> <a href="#">hq15Mm3L</a>	<input type="checkbox"/> <a href="#">intronEst</a>
<input type="checkbox"/> <a href="#">mouseChain</a>	<input type="checkbox"/> <a href="#">mouseChainLink</a>	<input checked="" type="checkbox"/> <a href="#">mrna</a>	<input type="checkbox"/> <a href="#">zoom1_hq15Mm3L</a>
<input type="checkbox"/> <a href="#">zoom2500_hq15Mm3L</a>	<input type="checkbox"/> <a href="#">zoom50_hq15Mm3L</a>	<input type="checkbox"/> <a href="#">cpqIsland</a>	<input type="checkbox"/> <a href="#">cytoBand</a>
<input type="checkbox"/> <a href="#">ensGene</a>	<input type="checkbox"/> <a href="#">estOrientInfo</a>	<input type="checkbox"/> <a href="#">firstEF</a>	<input type="checkbox"/> <a href="#">fishClones</a>
<input type="checkbox"/> <a href="#">fosEndPairs</a>	<input type="checkbox"/> <a href="#">gcPercent</a>	<input type="checkbox"/> <a href="#">geneid</a>	<input type="checkbox"/> <a href="#">genomicSuperDups</a>
<input type="checkbox"/> <a href="#">genscan</a>	<input type="checkbox"/> <a href="#">genscanSubopt</a>	<input type="checkbox"/> <a href="#">haplotype</a>	<input type="checkbox"/> <a href="#">knownCanonical</a>
<input checked="" type="checkbox"/> <a href="#">knownGene</a>	<input type="checkbox"/> <a href="#">mqcFullMrna</a>	<input type="checkbox"/> <a href="#">mqcGenes</a>	<input type="checkbox"/> <a href="#">mouseNet</a>
<input type="checkbox"/> <a href="#">mrnaOrientInfo</a>	<input type="checkbox"/> <a href="#">multizMm3Rn2</a>	<input type="checkbox"/> <a href="#">nci60</a>	<input type="checkbox"/> <a href="#">perlegen</a>
<input type="checkbox"/> <a href="#">recombRate</a>	<input type="checkbox"/> <a href="#">refGene</a>	<input type="checkbox"/> <a href="#">refSeqAli</a>	<input type="checkbox"/> <a href="#">rnaCluster</a>
<input type="checkbox"/> <a href="#">sgpGene</a>	<input type="checkbox"/> <a href="#">snpNih</a>	<input type="checkbox"/> <a href="#">snpTsc</a>	<input type="checkbox"/> <a href="#">softberryGene</a>
<input type="checkbox"/> <a href="#">stsMap</a>	<input type="checkbox"/> <a href="#">syntenyMouse</a>	<input type="checkbox"/> <a href="#">syntenyRat</a>	<input type="checkbox"/> <a href="#">twinscan</a>
<input type="checkbox"/> <a href="#">uniGene_2</a>	<input type="checkbox"/> <a href="#">vegaGene</a>	<input type="checkbox"/> <a href="#">vegaPseudoGene</a>	<input type="checkbox"/> <a href="#">xenoEst</a>
<input type="checkbox"/> <a href="#">xenoMrna</a>			

Plot width:

**Figure 5.3-11 Genome Annotation Features for Annotation Plots**

All annotation features included in the plot are hot linked to a corresponding UCSC annotation description page, which provides more detailed information about the annotations. The genome annotations provided by the webCGH viewer are obtained via the caBIO objects' Java interface to the distributed annotation system (DAS) server at UC Santa Cruz.<sup>27</sup>

<sup>27</sup> Details on the caBIO Java API to the UCSC DAS server (<http://genome.ucsc.edu>) can be found in Chapter 9 of The NCICB Technical Guide.

Additional control over how these annotation features will be displayed can be obtained by clicking on the hyperlinked text associated with each checkbox to connect to the genome server at UCSC. For example, clicking on [mrna](#) generates the new window shown in Figure 5.3-12. In addition to describing what that feature represents and how it is generated, the window allows features to be included or excluded using selection criteria such as author, library, tissue, etc.

**Human mRNAs from Genbank**

Display mode:

Filter: ☒ red ☐ green ☐ blue ☐ exclude ☐ include Combination Logic: ☒ and ☐ or

author:  library:  tissue:  cell:

keyword:  gene:  product:  description:

---

**Description**

The Human mRNA track shows alignments between human mRNAs in Genbank and the genome. Aligning regions (usually exons) are shown as black boxes connected by lines for gaps (spliced out introns usually). In full display, arrows on the introns indicate the direction of transcription.

**Method**

Genbank human mRNAs are aligned against the genome using the [blat](#) program. When a single mRNA aligns in multiple places, the alignment having the highest base identity is found. Only alignments that have a base identity level within 1% of the best are kept. Alignments must also have at least 95% base identity to be kept.

**Using the Filter**

The track filter can be used to change the color or include/exclude a subset of individual items within a track. This is helpful when many items are shown in the track display, especially when only some are relevant to the current task. To use the filter:

1. Enter a value in one or more of the text boxes to filter the mRNA display. For example, to apply the filter to all liver mRNAs, type "liver" in the tissue box.

**Figure 5.3-12 Filters and Annotation Feature Documentation From the UCSC Gene Server**

Because of the large number of annotations, it is recommended that the user first zoom in on a very defined segment of the chromosome before displaying any annotations. The annotations are displayed in the form of ‘tracks’ or rows under the graph.

As an example of how these features might be used, consider the region of chromosome 8 showing increased DNA copy number in the preceding examples. Multiple instances of elevated DNA copy number are observed there, and an obvious question becomes which genes map to this genome region. To answer this question, we select the checkboxes labeled **knownGene** and **mRNA** in Figure 5.3-12. The resulting annotation plot is shown in Figure 5.3-13.

In Figure 5.3-13, DNA copy number measurements are graphically integrated with the annotated genome features, which are shown as a set of tracks. Each track occupies a row in the plot and is associated with one particular annotation type (e.g. known genes), and each feature in a given track is represented by a rectangular box. The endpoints of the box correspond to the

endpoints of that feature within the chromosome. The genome position (in base pairs) is shown at the top of the plot.

The DNA copy number measurements are shown in the topmost tracks as solid bands of color representing the individual probes on the array. These colors correspond to the observed copy number change, with red indicating an increase and blue a decrease of the DNA copy number in the sample compared to the reference DNA. A color scale is displayed beneath the plot.

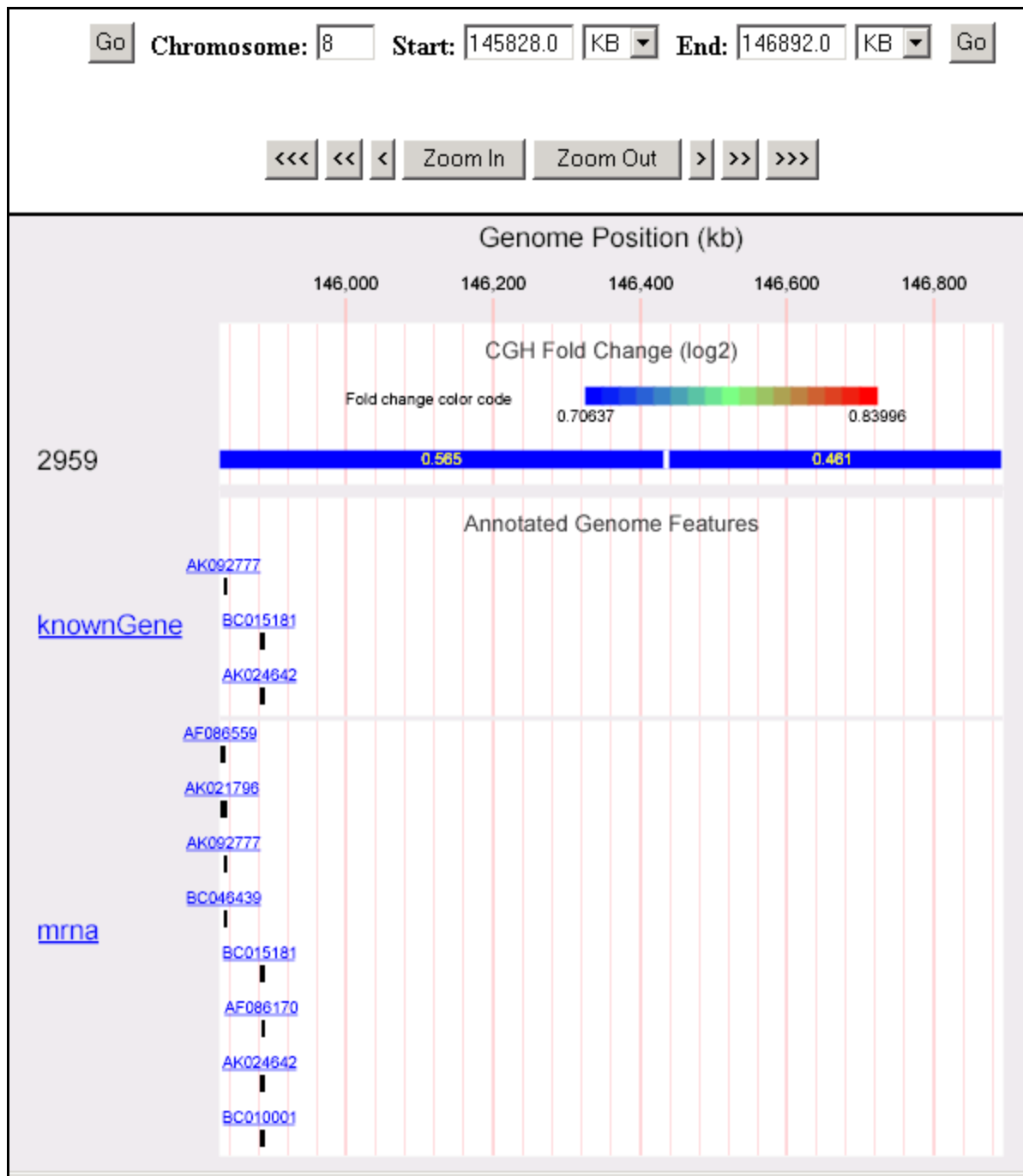


Figure 5.3-13 Annotation Plot for Chromosome 8

## **Animal Models and Cancer Images Analysis**



## 6.0 The Cancer Models Database

The Mouse Models of Human Cancers Consortium is a collaborative program designed to derive and characterize mouse models, and to generate resources, information, and innovative approaches to the application of these models in cancer research. Animal models that mimic the course of human cancers provide critical insight to the molecular etiology of the associated disease processes.

To date, the biological and genetic studies of murine strains have been extended to many other applications, including but not limited to screening targets for therapy, testing molecularly targeted agents for tumor prevention, establishing methods for early detection, and utilizing imaging technologies to detect malignant lesions and to monitor response to therapy.

The MMHCC's goal is to make information and materials concerning animal models of human cancer as widely available as possible to the entire cancer research community. In order to achieve this goal the MMHCC has initiated the development of three web-based resources:

- The [Emice](#) web site
- The [Cancer Models Database](#)
- The [Cancer Images Database](#)

The emice web site describes mouse models for human cancer, providing background information on human cancer and a general overview of cancer incidence, disease etiology, current methods for diagnosis and treatment, molecular alterations, and existing rodent models of human cancer listed by organ site. Furthermore, this web site contains links to recent publications of mouse models, offers learning and communication tools, catalogs research resources, and provides information about the MMHCC, its organization and activities.

The caMOD database contains information about animal models contributed by the broader research community, including the Consortium members. Many of the transgenic or knockout strains are available from the [MMHCC Repository](#) at NCI-Frederick, from the [Jackson Laboratory](#) in Maine, or directly from the principal investigators.

This chapter describes the Cancer Models Database at NCI and provides detailed instructions for using the web interface to the database. Chapter 7 provides a discussion of the caIMAGE database and its interfaces. The accompanying caCORE 2.0 Technical Guide documents the caCORE application programming interface to the caMOD database.

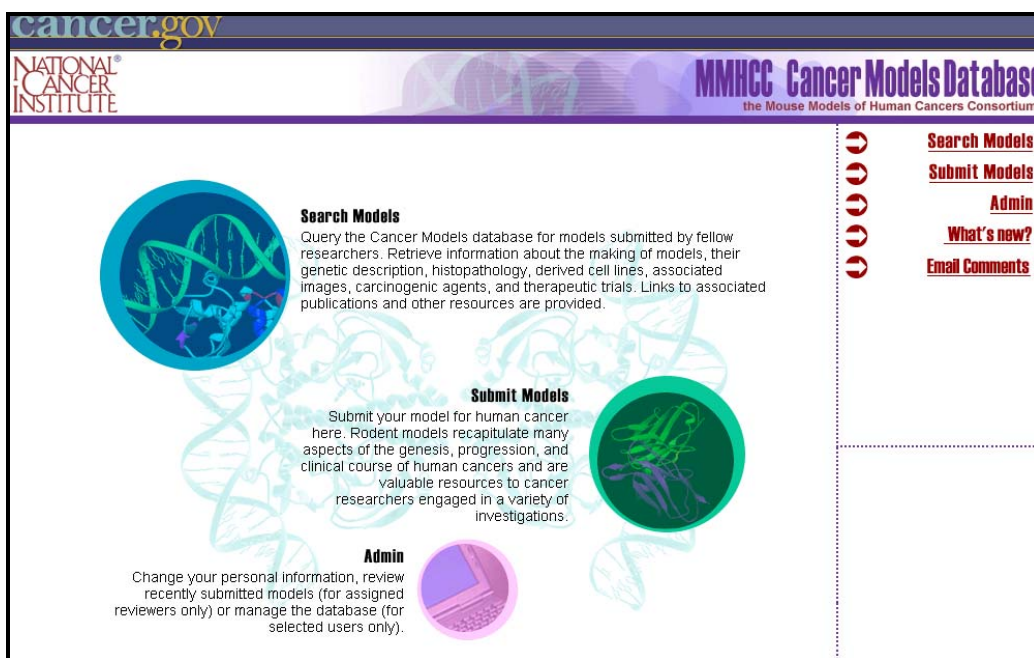
The caMOD database (<http://cancermodels.nci.nih.gov>) houses data for many cancer models, such as gastrointestinal, prostate, skin, and lung cancers in animals including mouse and rat. The data include histopathologic, genetic, expression profiling, and other biological data concerning the animal cancer models and their use for therapy or prevention studies. The database allows investigators to search publicly available data using search terms such as the model name, principal investigator's name, animal species, and strains, and provides contributors with personalized accounts for the submission of new animal models to the database.

### 6.1 Searching the caMOD Database

caMOD provides both simple and advanced search forms for generating database queries. The types of information a user can retrieve include histopathologies, genetic descriptions, derived cell lines, carcinogenic agents, therapeutic trials, model generation details, associated images,

and links to related publications. To access caMOD, begin by pointing your browser to the caMOD home page at <http://cancermodels.nci.nih.gov/mmhcc/index.jsp>.

The first screen that is shown posts a legal notice defining the terms of agreement for access to the database, and provides a link to the [Use Guidelines](#). First-time users are advised to read these documents before actually entering the site. Clicking on the link stating that you agree to these terms will then redirect your browser to the caMOD main page shown in Figure 6.1-1.



**Figure 6.1-1 The MMHCC Cancer Models Database**

The main page presents the user with three options for interacting with the database: search, data submission, and administration of user accounts. To begin a search, click on either the “Search Models” icon in the left panel or the textual key of the same name in the right panel. In response, a simple query form (Figure 6.1-2) will be presented where you can enter search criteria for:

- **Model Descriptor/Name:** This is simply the name of the model. The search will be restricted to all models *containing* the word or letter combination you enter.
- **PI's Name:** This is the name of the principal investigator who submitted the model. A pull-down list of contributing PIs is provided for you to select from.
- **Site of Lesion/Tumor:** This slot allows you to select models according to the specific sites that develop lesions. The animal model site locations must be specified using terms from the NCI's Enterprise Vocabulary Services. Thus, instead of typing directly into this textbox, you must select a term from the EVS DataTree that will be generated when you click the **Select** button.
- **Species:** Each model is associated with a particular species; a pull-down list of the available species is provided for you to select from.

As a simple example, this section demonstrates a search for mouse models involving the prostate glands with the string “mutant” contained in the descriptor. We begin by entering the string “mutant” in the model descriptor field, and press the **Select** button to select the lesion/tumor site. Figure 6.1-2 shows the basic query form for this example, after the site has been selected to be “Prostrate Glands,” and the species has been set to “Mouse (Mus Musculus).”

**Figure 6.1-2 An Example using the basic search query form**

Figure 6.1-3 shows the EVS DataTree for selecting the lesion/tumor site. The values in the lower left panel reflect the information displayed after the user has entered “prostate” in the text search box (upper left panel). The view of the EVS DataTree in the right panel is generated in response to selecting the **SHOW in DATATREE** option alongside the term “Prostate Glands.” The DataTree has been expanded under Reproductive System → Male Reproductive System → Prostate Glands, with the last term now highlighted.

**Figure 6.1-3 The EVS DataTree**

Clicking on a term in either the results list or the DataTree will cause a pop-up dialog to appear asking you to confirm this choice. Pressing **OK** will close the EVS window and return you to the search query form with the selected term appearing in the designated slot. Pressing **Cancel** instead of confirming the selection will allow you to continue browsing the tree for more specific terms.

Many terms in the DataTree having more specific “sub-terms” associated with them have folder-like icons to the left of the term. Clicking on a folder icon labeled with a “+” sign expands that term; clicking on an expanded folder icon (labeled with a “–” sign) will “collapse” that term and return the tree to its previous state.

For users who prefer to browse the DataTree directly, it is not necessary to use the left-hand search panel at all. Using the folder icons to expand and/or collapse subtrees can be applied to expose the desired term, which can then be selected directly by clicking on the hyperlinked text for that term.

To continue this example, select “prostate glands” by clicking on the hyperlinked name in the search results list. The name “prostate glands” will be inserted in the Organ/Tissue field on the submission page and the window with the vocabulary tree will close automatically.

Once you have filled in all of the desired search criteria in the basic search query form, clicking on the **Search** button will submit your query to the database. Figure 6.1-2 shows the partially filled in form specifying that mouse models involving the prostate glands with the string “mutant” contained in the descriptor should be retrieved.

It is *not* necessary to provide a value for each slot in the form—Figure 6.1-4 shows the results obtained by submitting the above form. Clicking on the **Model Descriptor** field will take you to the corresponding Information Pages for that model.

Search Results			
Your search returned 1 matching records.			1-1
No	Model Descriptor	Species	Tumor Sites
1	<a href="#">Nkx3.1 mutant mice</a>	Mouse (Mus Musculus)	Prostate Glands
Your search returned 1 matching records.			1-1

Figure 6.1-4 Results returned from a simple query

### 6.1.1 Using the Advanced Search Query Form

The Advanced Search Mode allows users to search the database with greater control over the search criteria. To enter this mode, click on the “Advanced Search Mode” hyperlink. The advanced search query form (Figure 6.1-5) that is generated includes the four fields that were previously defined in the simple form (Figure 6.1-2), plus the following additional fields:

- **Diagnosis:** Like the **Site of Lesion/Tumor** field, the allowed terms for this field are selectable from an EVS DataTree. Click the **Select** button to enter the search criteria for diagnosis.

Model  
Descriptor /  
Name

PI's Name

Site of  
Lesion /  
Tumor

Diagnosis

Species

[Simple Search Mode](#)

**Genetic Description:**

Gene Name:

☐ Engineered Transgene
☐ Targeted Modification

Genomic Segment  
Designator:

Select Inducing  
Agent for Induced  
Mutation:

**Carcinogenic Agents:**

Models with  
Carcinogenic  
Interventions:
☐ Check here to search for models with Carcinogenic interventions data

Select Chemical /  
Drug:

Select Growth  
Factor:

Select Hormone:

Select Radiation:

Select Virus:

Select Surgery:

**Phenotype:**

Phenotype:

**Cell Lines**

Cell Line:

**Therapeutic Approaches:**

Compound/ Drug:

Models with  
Therapeutic  
Approaches:
☐ Check here to search for models with therapeutic approaches data

**Micorarray Data**

Micorarray Data:
☐ Check here to search for models with micorarray data

**Figure 6.1-5 The Advanced Search query form**

- **Genetic Description:** This field allows you to select models according to the genes that were used to develop the model and the way in which they were manipulated. The



genetic description includes a gene name, a genomic segment designator, an inducing agent (for induced mutations), and a specification of how the gene was used—as a targeted mutation or engineered transgene. Any of these fields can be left unspecified.

- **Carcinogenic Agents:** Six categories of carcinogenic agent are defined: Chemical / Drug, Growth Factor, Hormone, Radiation, Virus, and Surgery. Each category has a pull-down list associated with it for you to select from.
- **Phenotype:** This field allows you to enter keywords for the model's phenotype.
- **Cell Lines:** Like the phenotype field, the cell lines specification need not specify the complete cell line name. Enter the name or parts of the name to search for a specific cell line.
- **Therapeutic Approaches:** Enter the name or parts of the name of the compound or drug that was used in therapeutic experiments. You can also use the checkbox to simply indicate that only models *with* therapeutic approaches data should be retrieved.
- **Microarray Data:** This field simply provides a checkbox to indicate that only those models associated with microarray data in the GEDP should be retrieved.

In using both the basic and advanced search query forms, the user should keep in mind that all of the fields containing values will be combined to form a conjunctive query. That is, *all* of the criteria must be satisfied by each of the models that will be returned. The interfaces do not at this time provide means for generating disjunctive queries stating that only some of the criteria must be satisfied. Specifying additional criteria will result in a more focused search. Specifying too many criteria, however, may omit relevant data due to an overly determined search.

## 6.1.2 Examining the Search Results

<a href="#">Back to Search Results</a> <a href="#">General Information</a> <a href="#">Genetic Description</a> <a href="#">Carcinogenic Interventions</a> <a href="#">Publications</a> <a href="#">Histopathology</a> <a href="#">Therapeutic Approaches</a> <a href="#">Cell Lines</a> <a href="#">Images</a> <a href="#">Microarrays</a>	
General Information for Model <b>BXH-2</b>	
<b>Model Descriptor</b>	BXH-2
<b>Species</b>	Mouse (Mus Musculus)
<b>Experimental Design</b>	BXH-2 mice will spontaneously develop acute myeloid leukemia when aged.
<b>Phenotype</b>	BXH-2 strain mice spontaneously develop acute myeloid leukemia (AML) due to chronic infection with a B-ecotropic murine leukemia virus (MuLV). Because the virus is passed from mother to offspring, researchers need not infect the mice themselves. Instead mice are simply aged until AML develops.
<b>Sex Distribution of the Phenotype</b>	Both Sexes
<b>Breeding Notes</b>	small and infrequent litters
<b>Submitted by</b>	<a href="#">Largaespada, David</a>
<b>Principal Investigator/Lab</b>	<a href="#">Largaespada</a>
<b>Available from Investigator</b>	For more information, please contact <a href="mailto:larga002@tc.umn.edu">larga002@tc.umn.edu</a>

Figure 6.1-6 General Information Page for a Retrieved Model

As described in the previous section, clicking on the **Model Descriptor** field in the Results page will take you to the Information pages associated with that model. Figure 6.1-6 shows the General Information page appearing on the top of a “stack” of related pages. These related pages contain other information associated with the model, including genetic description, carcinogenic interventions, publications, histopathology, therapeutic approaches, cell lines, images, and microarray data. The additional pages can be viewed by selecting the corresponding folder tabs running along the top of the current page.

## 6.2 Data Submissions

Users who have created personal accounts are encouraged to submit their models to the caMOD database. Although all models submitted are reviewed prior to being made public, the ability to create an account and submit a model is open to the public.

Clicking on the **Submit Models** button on the homepage (Figure 6.1-1) will take you to a Login screen where previous users can access their existing accounts and new users can create new accounts. To create an account, click on the “sign up now” hypertext link. This will open up a User Information page with the required fields highlighted in red. Note that the user’s name need not match the principal investigator’s name on this form. Upon submitting this registration form, the user is asked to select a username and password and, subsequently, is returned to the original Login page where these can now be entered to access the newly created account.

The Welcome page that is first seen when accessing a user account is customized for each user. New users will find basic instructions and guidelines for submitting models, along with links to a “Guided Tour” and more general help documents (Figure 6.2-1). First-time users can submit models by clicking on the **Submit Model** button at the bottom of the page.

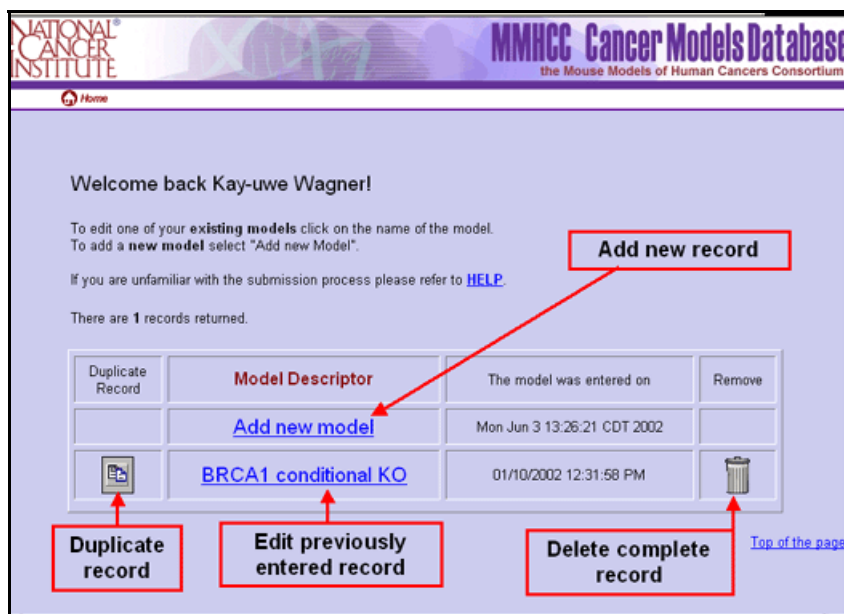


Figure 6.2-1 Welcome Page for New User

Established users who have previously submitted animal models will see a table listing these models on their Welcome page (Figure 6.2-2). The list itself, as well as the models it contains,



can be edited by clicking on the descriptor names. Clicking on the *Remove* icon (a trashcan in the rightmost field) for a particular model will result in that model being deleted from the list and from the database.<sup>28</sup>



**Figure 6.2-2 Welcome Page for a Returning User**

The *Duplicate Record* feature in the leftmost field can be used to speed up and simplify the submission of new models that are closely related to previously submitted ones. Selecting the *Duplicate Record* icon for an existing model will result in the insertion of a copy of that model to the list. This copy can then be renamed and edited to define a new model, as all existing models can be edited by clicking on their descriptor fields. Finally, it is also possible to add a completely new model by clicking on the first entry in the table labeled “Add new model.”

The process for submitting a new model commences with a General Information page providing an overview of the experimental design, animal phenotype, availability of the model, and so forth. This page must be completed before proceeding to the next step.

The required fields on the General Information page include:

- **Model Descriptor** – the commonly used name for the model. In some cases the model name might be the same as the genotype (e.g. Wap-Cre). The interface accepts multiple models with the same name, as records are stored via unique numerical identifiers, not by name.
- **PI's Email Address** – the email address of the principal investigator; if none is available, you must explicitly enter “none.”
- **Species** – the species of the model must be selected from the provided pull-down list. Species not included in this list can be requested for inclusion by contacting the NCICB [help desk](mailto:ncicb@pop.nci.nih.gov) at: [ncicb@pop.nci.nih.gov](mailto:ncicb@pop.nci.nih.gov).

<sup>28</sup> This action will result in permanent removal of the model, as the caMOD database does not archive deleted items.

- **Phenotype** – a general description of the phenotype, for example: “Expression of the SV40 early region in the mammary and prostate epithelium leading to invasive carcinoma development over a predictable time course. The transgene is expressed in mammary ductal cells and the terminal ductal lobular unit in virgin animals and without hormone stimulation.”
- **Record Release Date** – this field provides two options: (1) the record will be shown *immediately*, and (2) the record will be shown *after* a specified date.

Optional fields on the General Information page include genotype, availability of the actual model, strain, experimental design, sex distribution, and breeding notes. Many of these have controlled values, with selections provided to the user via pull-down menus and radio buttons. The free text type-in entries have extensive help documentation, which is available via the Help buttons provided alongside these entries.

The standard nomenclature for animal models can be both arcane and complex, and it is highly recommended that users unfamiliar with these terms consult the extensive notes available here. Topics included in the notes on nomenclature include

- Genetically engineered mice
- Chemically induced and targeted mutations
- Mice with spontaneous mutations
- Cloned mutations
- Coat color markers
- Congenic strains

In addition to the information provided directly on the help page, useful links to other resources providing further information, such as [Helpful Hints for Understanding Strain Nomenclature](#) are also available.

Following submission of the General Information page, the Continuing Submission Process page is generated, which contains links to all of the other forms you will need to complete the process. Because there are numerous ways to generate an animal model, there is no pre-ordained sequence of steps to follow, and the submission process is designed to gather information in ways that are most convenient to the individual users. The only requirement is that you first complete the General Information page and subsequently visit and fill out the remaining categories of information in whatever order is most suitable.

The remaining categories are those shown previously in Figure 6.1-5 as tabs running across the top of the *General Information* page. These categories are: genetic description, carcinogenic interventions, publications, histopathology, therapeutic approaches, cell lines, images and microarray data. The user needs only to visit and provide information for those pages that are relevant to the model being submitted. Thus, if there are no microarray data associated with the model, for example, then it is not necessary to provide any further information for that category.

The submission forms for each category of information can be accessed by selecting the corresponding entries in the leftmost column of the table provided on the Continuing Submission Process page (see Figure 6.2-3). Upon completion and submission of the form for a selected category, the user is returned to the Continuing Submission Process page, where a new link to the data now appears under that category. Clicking on this link returns the user to the associated submission form, where the information can be edited if necessary.

It is also possible to register multiple submissions within a single category. For example, a model involving a cross between a knock-out animal and a double transgenic animal, with ultraviolet radiation treatment applied, requires the submission of at least four additional forms beyond the General Information page. The additional forms include three Genetic Description forms (one for each transgene, plus one for the knock-out gene) and one Carcinogenic Intervention form for the UV-light treatment.

Figure 6.2-3 The Continuing Submission Process Page

For each submission, the user simply selects the appropriate [Enter...](#) link and a new blank submission page for that category will be generated. The entire session can be concluded (and resumed later if desired) at any point by clicking on the [Exit](#) button. All entered information will be temporarily saved but not subjected to the review process until the user designates that the model is complete.

Excellent examples of the submission process along with step-by-step directions for each category are available in the [Guided Tour](#).

### 6.3 The Admin Tool

The *Admin* tool on the home page allows users to manage the account information that was entered when they originally applied for an initial user account. The User Profile can be edited at any time, and includes the user's name, address, phone, and email address, along with the name of the principal investigator associated with that user. The additional administrative capabilities allow assigned reviewers to screen and review models and are password-protected.

## 7.0 THE CANCER IMAGES DATABASE

This chapter describes the caIMAGE database at NCI, and provides detailed instructions for using its web interface. The Cancer Images Database is being developed to host images of human and rodent cancer and normal tissue that are submitted by researchers.

Users can retrieve images and image annotations including species, tissue, gender, diagnosis, and image dimensions. The interface allows investigators to search the available data using terms such as the image name, principal investigator, animal species, and strains, and provides contributors with personalized accounts for the submission of new images.

### 7.1 Searching the caIMAGE Database

caIMAGE provides both simple and advanced search forms for generating database queries. To access caIMAGE, begin by pointing your browser to the caIMAGE home page, at <http://cancerimages.nci.nih.gov/>.

First-time users are advised to read the legal notice and [Use Guidelines](#) before actually entering the site. Clicking on the link stating that you agree to these terms will redirect your browser to the caIMAGE home page, shown below.

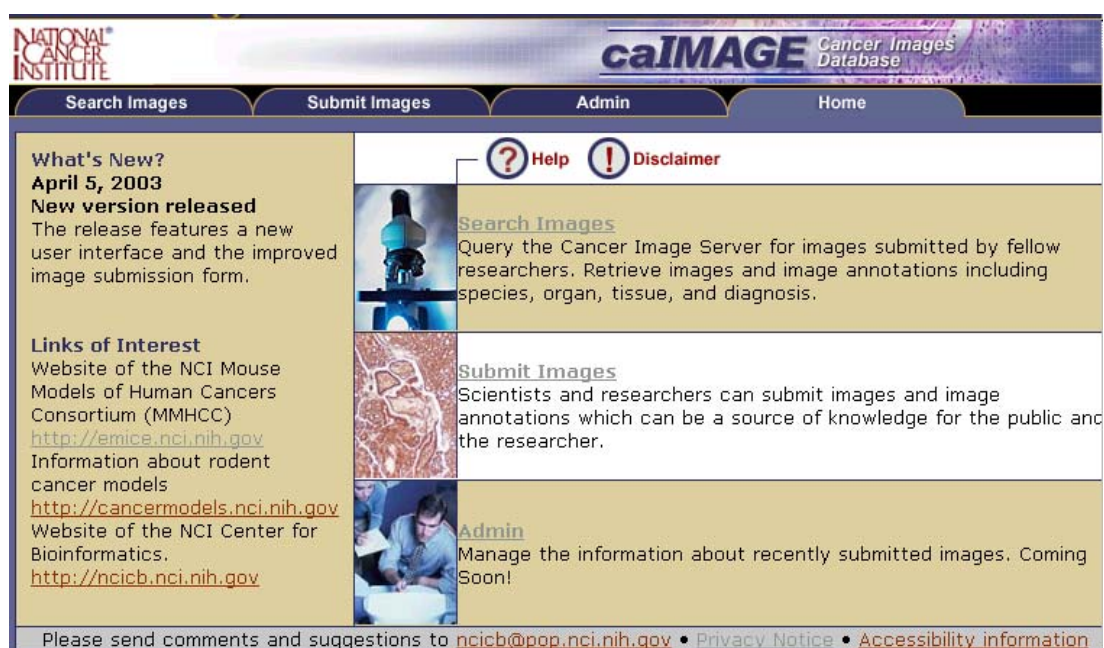


Figure 7.1-1 The caIMAGE home page

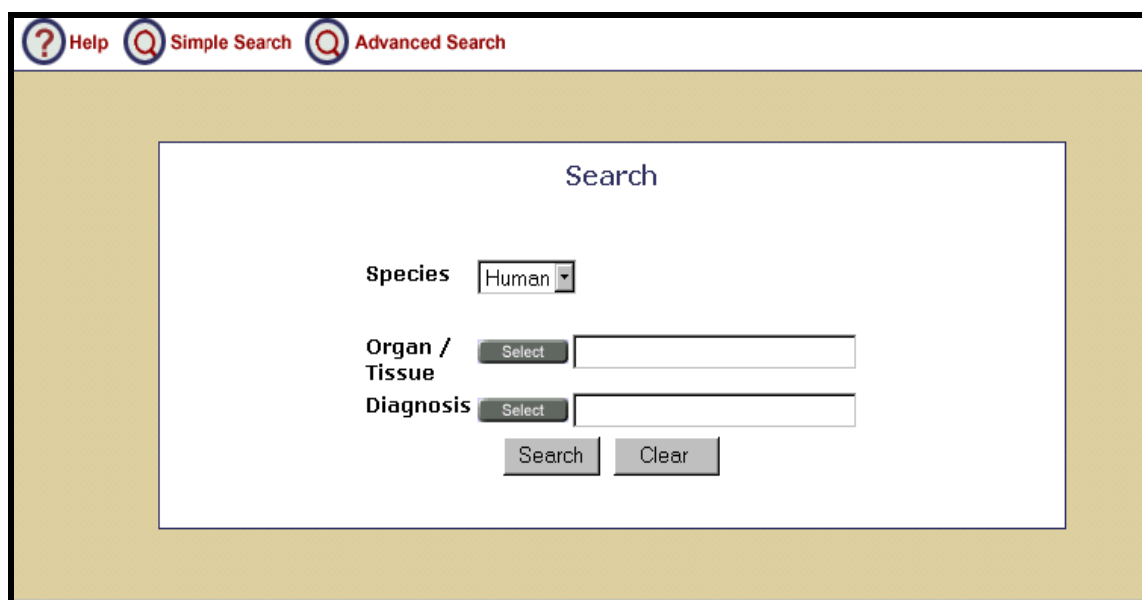
Most of the caIMAGE pages display a left sidebar and a central work area to the right of the sidebar, as in Figure 7.1-1.<sup>29</sup> The sidebar panel posts announcements about upcoming events and provides links to related sites. The right panel defines a central work area where various query and submission forms and result sets are displayed. Clicking on one of the folder tabs running across the top of the screen brings that “folder” to the front of the display. Each of the caIMAGE folder screens uses the same layout and sidebar panel—only the central work area in the right

<sup>29</sup> In many of the screen shots that follow, both the sidebar and caIMAGE banner have been cropped to improve the readability of the central work area.

panel is swapped out. The home page's work area summarizes the three main task groups that are accessed via the tab buttons, and provides hyperlinks to those services as well.

To begin a search, click on either the *Search Images* tab or the hyperlink by that name in the right panel. In response, a simple query form will be presented in the central work area (Figure 7.1-2) where you can enter search criteria for:

- **Species:** Each image is associated with a particular species; a pull-down list of the available species is provided for you to select from.
- **Organ/Tissue:** This slot allows you to select images by organ- or tissue-specific site locations, which must be specified using terms from the Enterprise Vocabulary Services. Thus, instead of typing directly into this textbox, you must select a term from the EVS DataTree by clicking the **Select** button.
- **Diagnosis:** This slot allows you to select images according to their associated diagnoses. Like the Organ/Tissue slot, values here must be selected from terms defined in the EVS DataTree, in this case, for diagnoses.



**Figure 7.1-2 The Simple Search Query Form**

The EVS Navigator window (Figure 7.1-3) is divided into a Search panel (left) and a DataTree panel (right). The top region of the Search panel provides a textbox for entering search terms, and the lower region is used to display search results. Since the EVS search engine performs both semantic and orthographic matching, the search results will include syntactic matches as well as alternative terms or aliases. Each match in the SEARCH: Results area has a **SHOW in DATATREE** icon next to it. Selecting that icon causes the associated branch of the DataTree to expand and be highlighted.

For example, to search for images related to human colon cancer, select “human” in the pull-down species menu and press the **Select** key next to the Organ/Tissue field. When the EVS Navigator window opens, type “colon” in the search field (as shown in Figure 7.1-3) and click the **Search** button to find related concepts in the EVS vocabularies.



To select a term for entry on the query form, you can click on the hyperlinked name in either the search results listing or the corresponding node in the DataTree. To see the node in the DataTree, click on the **SHOW in DATATREE** icon next to it.

Clicking on a term in either the results list or the DataTree will cause a pop-up dialog to appear asking you to confirm this choice. Pressing **OK** will close the EVS window and return you to the search query form with the selected term in the designated slot.

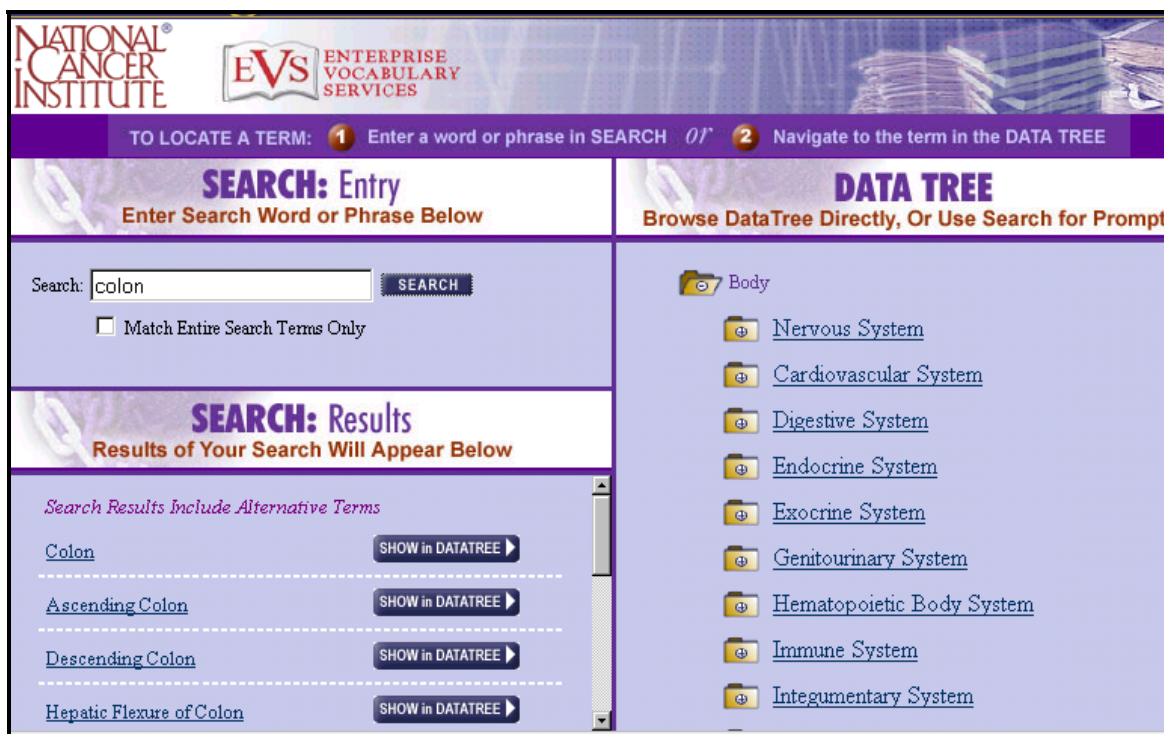


Figure 7.1-3 The EVS Navigator Window

Pressing **Cancel** instead of confirming the selection will allow you to continue browsing the tree for more specific terms. All terms in the DataTree having more specific “sub-terms” will have folder-like icons to the left of the term. Clicking on a folder icon labeled with a “+” sign expands that term; clicking on an expanded folder icon (labeled with a “-” sign) will “collapse” that term and return the tree to its previous state.

To continue this example, select “colon” by clicking on the hyperlinked name in the search results list. The name “colon” will be inserted in the Organ/Tissue field on the submission page and the EVS window will close automatically.

At this point you may choose to further constrain the search by also specifying a diagnosis. Alternatively, you can execute the search immediately by pressing the **Search** key. In some cases, leaving the diagnosis unspecified may produce too many results, so you will want to add this additional constraint. In other cases, specifying a diagnosis may over-constrain the search and produce too few results. Note, however, that it is not possible to specify a diagnosis without first specifying the organ/tissue, as the diagnoses are in general site-specific.

For this example go ahead and execute the search, leaving the diagnosis field unspecified. Figure 7.1-4 shows the first three of the eight results that were found. As shown, each result is represented by a postage stamp-sized image and a record just to the right of that image. Each record contains five fields specifying a description of the image, an organ site, a species name, the name of the investigator who submitted the data, and the dimensions and magnification of the image. Some records may contain additional information; for example, the third result includes information on the staining that was applied.

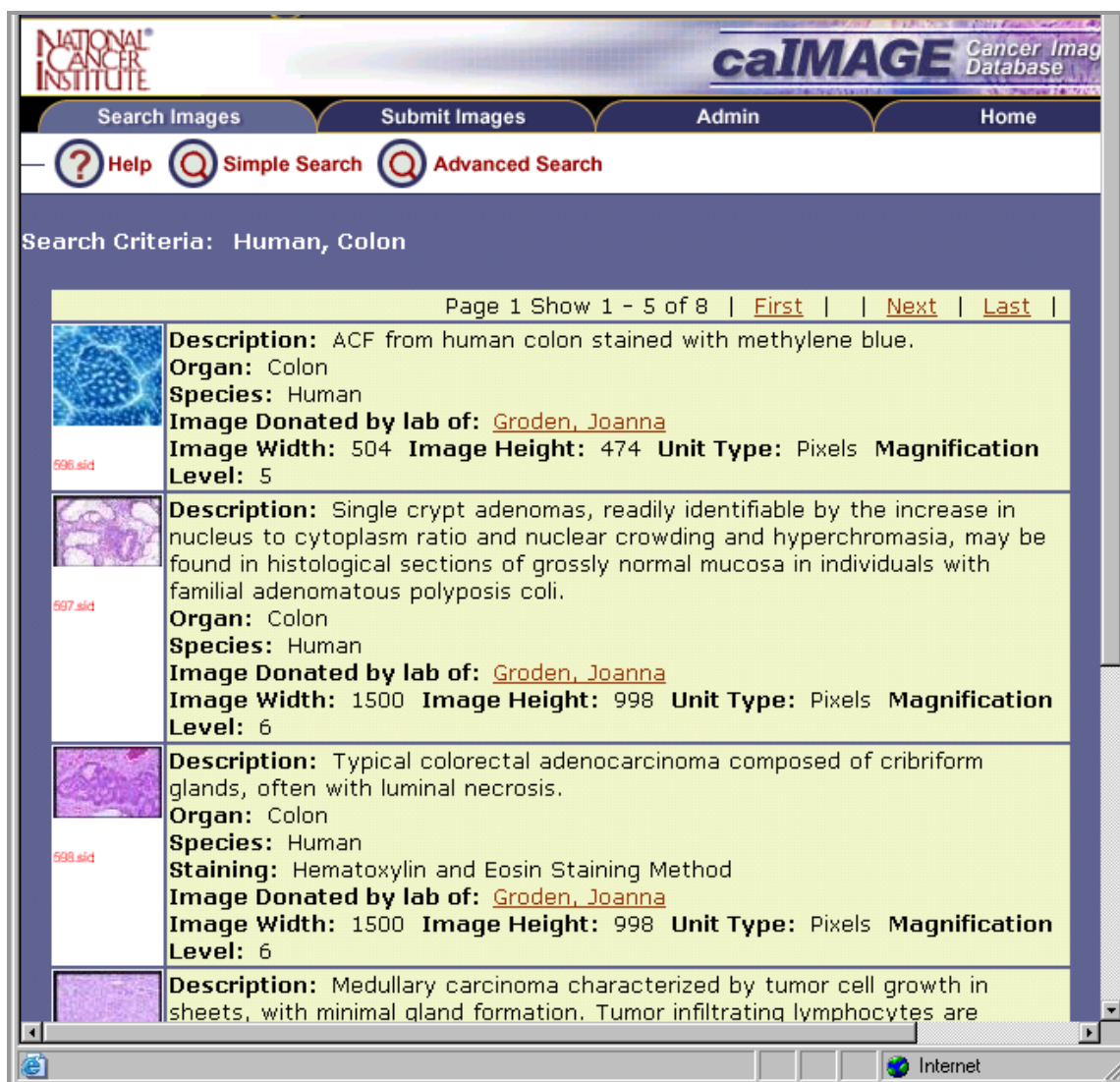
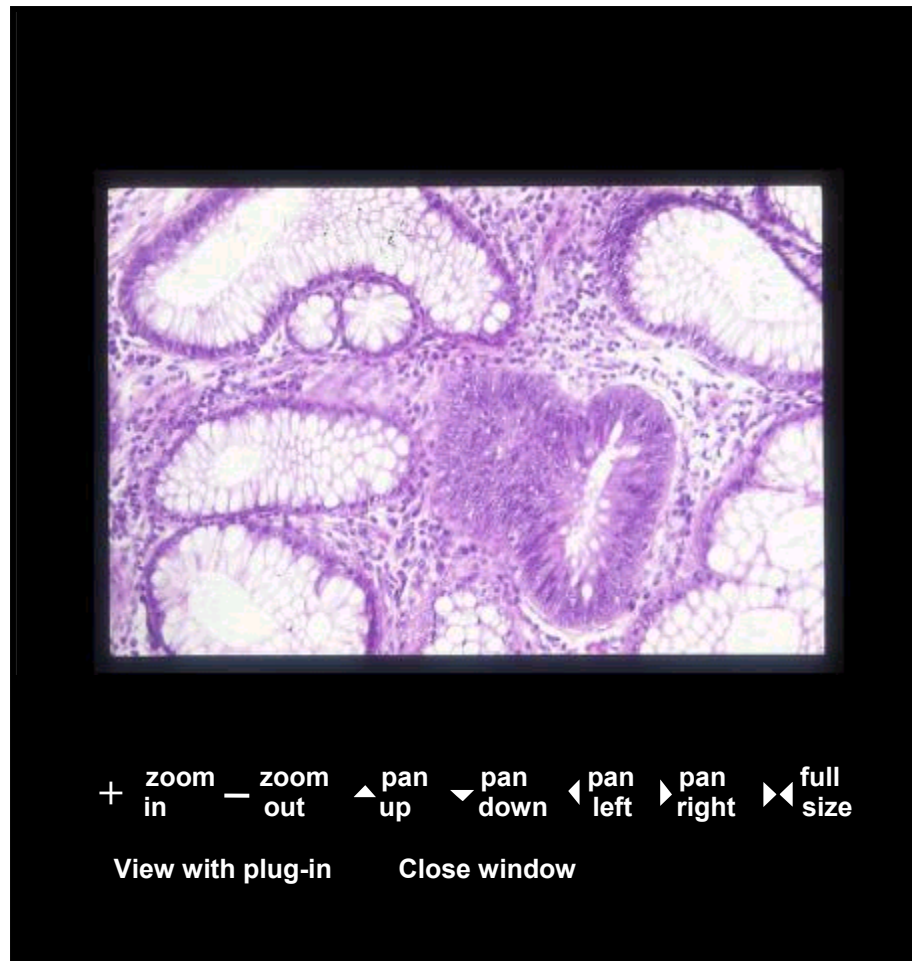


Figure 7.1-4 Search Results for Human Colon Cancer Images

Clicking on the investigator's name will open an email message addressed to that investigator, which you may use to request additional information if needed. Clicking on the image will open a new window where you can explore the image in more detail. The viewer shown in Figure 7.1-5 provides simple controls for panning up/down and right/left, and for zooming in and out.





**Figure 7.1-5 Exploring an Image with the Viewer**

The *Advanced Search* interface is somewhat more complicated to use but provides greater control over the search. To enter the Advanced Search mode, click on the hyperlink by that name (or its associated magnifying glass icon) at the top of the Simple Search interface.

Figure 7.1-6 shows the advanced search query form that appears in the central work area. In addition to the Species, Organ/Tissue, and Diagnosis fields that were used on the Simple Search form, six new fields are introduced:

- **Image Name/Number:** This constraint should only be used when you know the exact name or number for the image. This option is provided for users who are not “searching” for images, but instead, wish to access images they have worked with previously.
- **Staining:** Approximately 40 different staining methods are listed in the pull-down menu provided for this slot. Some, but not all records include this information. Applying this constraint to your search may exclude certain records due to missing information rather than explicit mismatches.

- **Gender:** The choices are male, female, and unknown. Note that choosing “male” or “female” will exclude all images where the gender is unknown, and similarly, specifying “unknown” will exclude all records where the gender information is available.
- **Mouse Strain:** Approximately 60 different mouse strains are listed in the pull-down menu.
- **PI's Name:** All PIs who have submitted images to the database are listed as choices in the pull-down menu.
- **Institute:** All institutes and organizations associated with PIs who have submitted images are listed in the pull-down menu.

The screenshot shows a web-based search interface. At the top, there are three tabs: 'Help' (with a question mark icon), 'Advanced Search' (with a magnifying glass icon), and 'Simple Search' (with a magnifying glass icon). The 'Advanced Search' tab is selected. Below the tabs is a yellow header bar. The main content area is titled 'Advanced Search' and contains several sections separated by blue bars. The 'Image' section has a text input for 'Image Name / Number', a dropdown for 'Staining', and a dropdown for 'Gender'. The 'Specimen' section has a dropdown for 'Species' (currently set to 'Mouse'), a 'Select' button and text input for 'Organ / Tissue', and another 'Select' button and text input for 'Diagnosis'. The 'Submitter' section has a dropdown for 'PI's Name' and a dropdown for 'Institute'. At the bottom, there are 'Search' and 'Clear' buttons.

Figure 7.1-6 The Advanced Search Query Form

The results screen for Advanced Search is identical to that shown for Simple Search. The Advanced Search query form is in itself no more complicated than the Simple Search query form; the complication lies in knowing when to apply additional constraints to your search and when to leave these unspecified.

## 7.2 Data Submissions

Users who have created personal accounts are encouraged to submit their images to the database. Clicking on the [Submit Images](#) hyperlink on the homepage (Figure 7.1-1)—or on the folder tab by that name (from any screen)—will take you to the login screen for data

submissions. While the entire user community is welcome to browse and search the database, only registered account holders may submit data.

The data submission login screen allows returning users to log in directly and provides new users with the opportunity to register for an account. To create an account, click on the “sign up now” hypertext link. This will open up a User Information page with the required fields highlighted in red.<sup>30</sup> Upon submitting this registration form, the user is asked to select a username and password, and is subsequently returned to the original login page where these can now be entered to access the newly created account.

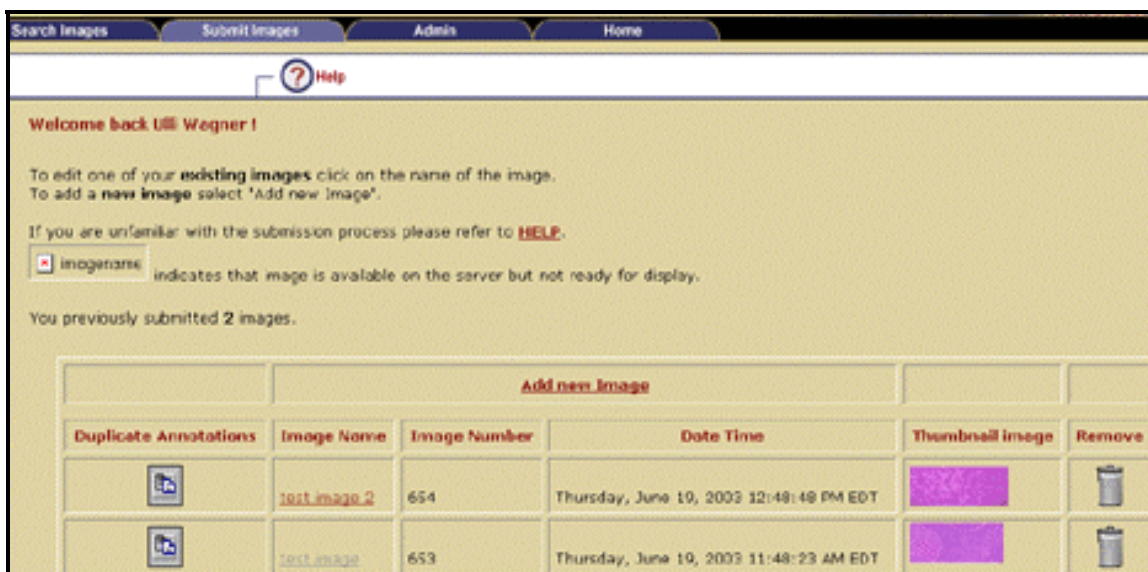


Figure 7.2-1 The Data Submission Welcome Screen for returning users

The Welcome page that is first seen when accessing a user account is customized for each user. Users who have previously submitted data will see a table listing these images on their Welcome Page (Figure 7.2-1). New users will see the same page displayed, but the table will of course, be empty. First-time users can submit data by clicking on **Add New Image**, as described below.

The list of submitted images, as well as the images it contains, can be edited by clicking on the image names. Clicking on the **Remove** icon (a trashcan in the rightmost field) for a particular image will result in that image being deleted from the list and from the database.<sup>31</sup>

The “annotations” for an image are defined by the values you provide on the image submission form shown in Figure 7.2-2. The **Duplicate Annotations** feature in the table of Figure 7.2-1 can be used to speed up and simplify the submission of new images that are closely related to previously submitted ones. Selecting this field for an existing image will result in the insertion of a new record whose annotations are copied from the existing image.

<sup>30</sup> Detailed instructions for creating a new account are included in the online [Guided Tour](#) for data submission.

<sup>31</sup> This action will result in permanent removal of the model, as the caIMAGE database does not archive deleted items.

The name of the new image will also be copied from the existing image, with “copy” appended to it. The image itself, however, will not be copied, and the **Thumbnail Image** field will post the warning “Image not found.” This newly copied record can then be edited by clicking on the **Image Name** field. Once all of the annotations and the image name have been modified to correspond to the new submission, the new image can be uploaded. All previously submitted images can likewise be edited at any time by clicking on the **Image Name** field.


Image Submission	
* required field	
Specimen	
<b>Species*</b>	<input type="text"/>
<b>Organ / Tissue*</b>	<input type="button" value="Select"/> <input type="text"/>
<b>Diagnosis</b>	<input type="button" value="Select"/> <input type="text"/>
<b>Gender</b>	<input type="text"/>
<b>Mouse Strain</b>	<input type="text"/>
<b>If Other Strain</b>	<input type="text"/>
<b>Promoter</b> (for genetically engineered Models only)	<input type="text"/>
<b>Gene</b> (for genetically engineered Models only)	<input type="text"/>
Publication	
<b>PMID</b> (PubMed Identifier)	<input type="text"/> 
Image	
<b>Image Modality</b>	<input type="text" value="Histology"/>
<b>Image Name*</b>	<input type="text"/>
<b>Image Description</b>	<input type="text"/>
<b>Staining</b>	<input type="text"/>
<b>If Other Staining</b>	<input type="text"/>
<b>Image Upload*</b>	<input type="text"/> <input type="button" value="Browse..."/>
<input type="button" value="Submit"/> <input type="button" value="Clear"/>	

Figure 7.2-2 The Submission Form for New Images

Adding a new image from scratch by clicking on **Add New Image** will bring up the submission form in Figure 7.2-2. All fields on that form that are highlighted in red must be filled out before actually uploading the image.

Information on the submission form falls into three broad categories: features of the specimen, publications associated with the image, and features of the image. Step-by-step instructions for filling out this form and uploading the image itself are provided in the online Guided Tour.